AN ANALYSIS OF TEST/RETEST EXPERIMENTS ON THE 1972, 1973, 1974, AND 1978 GENERAL SOCIAL SURVEYS

GSS Methodological Report No. 8

December, 1979

Tom W. Smith
and
C. Bruce Stephenson

This research was done for the General Social Survey project. The project is under the direction of James A. Davis and is supported by the National Science Foundation, SOC77-03279.

In a test/retest or panel design, an individual is administered the same stimulus (question) at two or more points in time. This repetition of measuring the same attribute of the same individual is used for two basic and distinct purposes: (1) to calculate measurement error, and (2) to measure true change. The fundamental problem with the test/retest design is that it measures these two features concurrently and, as we shall see, it is at best difficult to disaggregate these two components of test/retest data and come up with separate and accurate measurements of error and change.

In this paper we will examine the test/retest design's ability to calculate measurement error and true change. In particular, we will inspect evidence from test/retest experiments on the 1972, 1973, 1974, and 1978 General Social Surveys--GSSs (see Appendix 1). After briefly considering the definition, nature, and source of measurement error and true change, we will examine the general adequacy of the test/retest design and compare its strengths and weaknesses to alternative methods of handling measurement error and true change. Next, we will start to unravel the components of test/retest consistency[1] by examining (1) several special cases in which assumptions can be made about the amount of measurement error and true change that is present, (2) various individual-level

[1]Note that this use of the word "consistency" is different from discussion of inconsistency errors and consistency coefficients in Cureton, 1968. We use "consistency" to cover the joint effects of both measurement error and true change. It can be thought of as the proportion giving the same response at both times (in discrete variables) or as the raw correlation between responses to a question at both times (in continuous variables).

techniques for separating error and change, and (3) various aggregate-level techniques such as the three (or more) wave techniques proposed by Heise, Wiley and Wiley, Henry, and others. Finally, we will consider the meaning of these findings on the application and use of test/retest designs and the calculation of measurement error and true change on typical sociological items.

In a test/retest situation, the same set of individuals are asked the same questions at two different times, Responses between the two surveys will neither remain the same or change because of (1) measurement error or (2) true change. If we assume that there is a single correct response to every question, then measurement error occurs whenever the correct answer does not appear in the final data set. The complement of measurement error is reliability. When there is no measurement error there is perfect reliability and as measurement error increases from zero, reliability decreases apace. True change occurs when a person's correct answer changes between the test and retest. The complement of true change is stability.[2] When there is no true change, there is perfect stability and as true change increases from zero, stability decreases correspondingly.

Measurement Error and Reliability

Among discrete variables, measurement error can be expressed as the occurrence of responses other than the true score. On the aggregate level, it's the proportion of cases in which the observed score does

---

[2]To keep our terminology as clear as possible, we note that our use of stability here and of stability coefficients later on refers to true change and not to measurement error. Their use should not be confused with the division of reliability coefficients into consistency and stability coefficients as in Cureton, 1968.

not equal the true score. Among continuous variables, it is the difference between the true score and the observed score. On the aggregate level, it is the difference between the total observed variance and the true score variance.

Reliability for discrete variables is the simple inverse of measurement error, the proportion of cases in which the observed score equals the true score. For continuous variables, it is the ratio of the true score variance to total variance.[3]

Classic test theory assumed uncorrelated error. Sometimes this is seen as empirically representing the actual nature of error, other times it is seen as merely the simplest representation of possible relationships between measurement error and true scores, and in still other occasions, measurement error is defined to exclude anything but random error with correlated error being considered as part of validity and not reliability. In this discussion, measurement error is considered to be made up of both random and nonrandom error. This formulation of measurement error makes the measurement of reliability and the disaggregation of test/retest

---

[3]Reliability is frequently defined in terms of consistency. Bohrnstedt (1969:83) observed, "What is meant by reliability? Perhaps the best synonym is consistency." Similarly, Nunnanlly (1975:311) noted, "Reliability concerns the extent to which measurements are repeatable . . ." and Selltiz (1976:161) defined reliability as "the extent to which measures give consistent results." We, however, find these definitions inadequate since they are all based on the assumption of random error and do not consider the repetition of false answers. Less seriously, they are inadequate because they down play true change (this qualifier is routinely recognized in later elaborations by these authors). The essence of reliability is accuracy or truthfulness. Consistency is an attribute of reliability only when measurement error is random (and even then random repetition of error must be figured in) and true change in nil. There is, thus, a conditional and not general association between reliability and consistency.

data into measurement error and true change components much more difficult. The simple classic formulation states that the observed score (x') equals the true score (x) plus measurement error (e) or

$$x' = x + e$$

Taking the case where measurement error is actually made up of random and nonrandom error and the nonrandom error is a linear function of the true score, we find that

$$e = c + dx + u$$

where u is random error. Substituting, we see that

$$x' = c + (1 + d)x + u$$

Thus the observed score equals the true score plus correlated error of the true score (d) plus a constant (c) plus random error (u). In terms of variances, we get

$$V_{x'} = (1 + d)^2 * V_x + V_u$$

The observed variance equals the true variance times correlated error plus the random error variance. While the inclusion of correlated error has the undesirable property of making the relationship between true and observed scores more complex and less determinable, it is nevertheless necessary since (a) correlated error is quite common and (b) the impact of correlated error is less predictable and more likely to distort analysis than uncorrelated error. Correlated errors of various kinds are typical in most measurement instruments and occasionally this type of error is quite large. Marginal distributions are influenced by random error in known ways. If the distribution is equal for all categories, the distribution will not be changed. If the measurement error is less than 50 percent, the observed distribution will underestimate the marginal skew

(e.g., the distribution will move towards a 50/50 split in a dichotomy). On continuous variables, means will be unchanged by random error. Nonrandom error, on the other hand, will always change the mean or the distribution of continuous and discrete variables and its direction is never known unless the operation of the nonrandom error is known (e.g., a social desirability effect will increase the proportion in the normative category or move the mean toward the normative end of a scale). Looking at interrelationships, we find that random error will always attenuate simple product moment correlations and coefficients of determination. Nonrandom error can, however, either attenuate or increase these statistics. In brief, nonrandom error can create a much less predictable pattern of distortion than random error and because it is also a common and occasionally large source of error, its impact must be considered in all discussions of measurement error and reliability.

Measurement error results from many sources. It is probably impossible to specify even generally all of the possible sources of measurement error. Nevertheless, it is useful to list the range and variety of measurement error sources.

In our scheme of the sources of measurement error, we have placed emphasis on where and why error is introduced.[4] We have specified three broad locations for measurement error: (1) the individual question itself; (2) other characteristics of the interview situation; and (3) post-interview occurrences. In the following section, we will discuss the major reasons that measurement error may be introduced at each of these points.

---

[4]Two alternative schemes for classifying measurement error can be found in Cronbach, 1970, p. 175, and Sudman and Bradburn, 1974, pp. 1-23.

In the case of the individual question, measurement error may come from either (a) the form or construction of the question or (b) its substance or topic. No question is perfect and some respondents and/or interviewers will misconstrue or misunderstand even the best of questions. Among the particular characteristics of question wording that may contribute to incorrect responses are

1. incoherence--a question that makes little or no sense perhaps because of the typographical omission of a key word string or a poor translation;

2. being double-barreled--a question that focuses on two distinct elements simultaneously (e.g., Do you approve of killing baby seals and using porpoises to hunt tuna?);

3. difficult or technical vocabulary or jargon;

4. being too involved (e.g., "Have you heard about or followed the case of Gary Mark Gilmore, the man convicted of murder in Utah who requested that he be executed by a firing squad, or haven't you heard about or followed this? (As you know,) Gilmore was convicted of murder and was sentenced to be executed. He asked the state of Utah to execute him immediately without any further appeals. Because of the publicity he received for asking for his own execution, Gilmore has received offers of sizable sums of money to publish his memoirs after he is dead. He wants some of this money to go to the families of the people he murdered. The Utah Supreme Court reviewed the case and backed up the state on executing Gilmore by firing squad. Lawyers for Gilmore's mother and the American Civil Liberties Union asked that the death sentence be put off, partly because, they claim, it is cruel and unusual punishment and partly because Gilmore's behavior had become a commerical proposition, involving large sums of money. All in all, do you think Gary Gilmore ought to be executed by a firing squad or not?");

5. being too vague or simplistic (e.g., "Do you support America's foreign policy?" or "Do you think something ought to be done about morality?");

6. inconsistent response categories--either inconsistent with the question or with other response categories;

7. indistinct response categories--difference between categories unclear (e.g., "How much do you like ice cream? Is it super, really great, kinda special, nice, so-so, not the best, or second-rate?");

8. incomplete or restricted response categories not covering all possible responses (e.g., "How often do you bathe? Daily or annually?" or "Do you favor or oppose the Roman Catholic Church's opposition to women priests? Yes, No, or Don't know." Responses such as "It's their business, not mine," or "I'm Lutheran, that's up to the Catholics" are not covered); and

9. screens and skip patterns (which increase dramatically the number of "no answers" for a question).

In brief, these and related factors will hinder the determination of the correct response.

Measurement error may also occur even when there is no misinterpretation or confusion. Even with a well-worded question that is clearly understood, respondents may give an incorrect response for a variety of reasons. First, they may lie to hide an unpopular attitude (a social desirability effect), conceal some personal information (e.g., gun ownership), or for some other reason (e.g., acquiescence). Second, they may understand the question but not comprehend the issue being addressed. Rather than honestly replying, "Don't know," they may give a disingenuous substantive response (perhaps to cover their ignorance on the matter). Third, they may have an uncertain, borderline, or ambiguous response to an item, but through some random process akin to Brownian motion, choose a particular response. For example, a person evenly divided between agreeing and disagreeing with a statement may flip a mental coin and say agree when the correct response should have been something like, "I'm undecided." Or, in response to the query, "What do you think is the ideal number of children for a family to have?" a respondent might reply "five or six." To pinpoint the coding, the interviewer might probe for an exact number. To accomodate, the respondent might again flip the mental coin and reply "five" even though "five or six" was the true and correct answer. Fourth, respondents may give a wrong answer because

what a respondent believes to be true is actually false.[5] Last, a respondent may give a wrong answer because of a miscalculation or temporary memory lapse (e.g., errors of omission and telescoping on recall questions or rounding on years of education). In sum, even when respondents understand a question, they may give an incorrect answer because of out-right lies, accomodations to simplify their response or its recording, misremembering, or false knowledge.

Next, measurement error can come from other attributes of the interview situation. These include (1) the form and method of administration, (2) the survey instrument, (3) the interviewer, and (4) the respondent. The basic forms of administration (telephone, face-to-face, and self-administered) each have certain strengths and weaknesses that affect the number and types of measurement error. For example, telephone interviewing may reduce distortion from the presence of others but may cause greater fatigue; face-to-face interviewing may increase general rapport but increase the response effects from race of interviewer-respondent interactions; and self-administration may reduce social desirability effects but increase error among the hard-of-seeing or semi-literate.

---

[5] Counting as measurement error a factually incorrect response, even though the respondent does not know that he is wrong, is consistent with our definition of measurement error but can lead to some troublesome ramifications. It would apparently mean that all incorrect answers on an aptitude or achievement test would be measurement error. In these cases, however, information is not being collected but knowledge is being measured. Measurement error occurs whenever the respondent does better or worse than his true ability. If it is assumed that a person either truely knows or does not know the answer to a particular computation, then measurement error occurs when someone who does not know the answer guesses or miscalculates to the right answer or when someone who knows the answer stumbles into the wrong answer. The issue becomes even murkier when something like a timed test is involved. In this instance, a respondent may know how to do all the computations but may not be able to do them all or do them all inerrantly in the time allowed. These complex cases tend to be psychometric rather than sociological.

Attributes of the survey instrument that can cause measurement error are placement, order and context effects, response set, and other contaminants.[6] Interviewer attributes that can affect measurement error are of two types: (1) the skill and ability to conduct an interview and (2) personal characteristics such as age, sex, race, and social class that may interact with similar characteristics of the respondent. Respondent attributes that can affect measurement error are parallel: (1) the willingness and ability of the respondent to fulfill the role model of respondent and (2) fixed, observable characteristics that may interact with those of the interviewer. In brief, there are a whole series of factors that can cause response error during the interview that are entirely separate from each question taken independently.

Nor does error stop at the doorstep as the interviewer leaves. Two basic types of error can occur after interview. The one is intentional distortion or data doctoring. This includes the submission of invalid ("made-up") interviews, slanted coding of open-ended questions, and wholesale falsification of the data. The other source of measurement error is transference error. In order to discuss this form of error, we will go back to the interview and briefly trace data transference from start to finish.

Tranference error can occur whenever there is a transmittance of information. Taking the typical case of the face-to-face personal interview, the following major steps occur at the time of the interview: (1) from written questionnaire to eye of interviewer, (2) from interviewer's

---

[6] Of course, the context induced response to a question may be the correct answer to the item given the additional stimuli provided by the prior question just as a response to a loaded question may be the correct response given the "loading." However, what one presumably wants is the correct answer to a question independent of the artificial impact of a context effect or loaded phrase.

eye through brain to mouth, (3) from interviewer's mouth to respondent's ear, (4) from respondent's ear through brain to mouth, (5) from respondent's mouth to interviewer's ear, (6) from interviewer's ear through brain to hand, and (7) from interviewer's hand to questionnaire. (In cases where a repetition or clarification is made or requested, the number of transferences increases greatly, although in general the net result of these additional exchanges of information will be to reduce measurement error rather than to increase it.) At each and every one of these steps, measurement error can occur. It can happen whenever there is a slip of the tongue, blink of an eye, skip of the brain, or jerk of the hand. Such spasmodic errors always occur to some degree but their frequency can be affected greatly by physical or mental impairments or particularities on the part of the interviewer, respondent, or both, such as a hearing loss, speech defect, senility, mental retardation, inebriation, accent, etc., and distractions such as telephone calls or crying babies. Beyond the interview stage, the major steps at which information is changed or transferred are (1) interviewer editing, (2) coding, (3) keypunching, (4) cleaning, (5) reformating and recoding, and (6) data duplication. Again, error will always occur at any one of these stages and their frequency will be greatly influenced by the quality/carefulness of the data processing personnel and procedures and the amount of double checking that is done (e.g., verification, data editing, call backs, etc.).[7]

---

[7]Of course, certain steps, in particular interviewer editing and cleaning, will generally reduce the absolute number of errors rather than increase them.

From the preceding discussion, it is clear that there are many possible sources of measurement error.[8] Some error is assoicated with individual questions or scales such as error from poor wording or response set while other error occurs largely independent of individual items such as keypunching or inept interviewing. The item-specific and generalized error can be either random or nonrandom although there is some tendency for item-specific error to be nonrandom and generalized error to be random. In brief, measurement error has many causes, can occur at various stages of data collection, and may be either random or nonrandom in form.

True Change and Stability

True change is the difference between real values at two time points. If the real values have not varied, then true change is nil. If all real values have altered, then true change is complete. For a dichotomy, true change is usually defined as the number in category 1 at time one minus the number in category 1 at time two divided by the number of cases in category 1 + 2 at time one. For a polychotomy, it

---

[8] It may seem that since there are so many opportunities for error that there is little room left for anything else. Several factors mitigate against that. First of all, the probability for each type of error is fairly small (although rarely precisely known). Second, the impact of multiple measurement error is not simply cumulative. If measurement errors are uncorrelated, then the amount of net measurement error that will result when multiple errors occur will be reduced by measurement errors that cancel each other out. (Net measurement error on a dichotomy occurs only when there are an odd number of errors. If there are an even number of errors, the net measurement error is nil.) For example, if two types of measurement error occur and each leads to 10 percent of the cases being switched from plus to minus, then the net percent of cases incorrectly measured will be 18 percent if the errors are uncorrelated. If the errors are positively correlated, then the net amount of error will be even less. If the errors are negatively correlated, then the amount of error will be higher. (If errors are perfectly negatively correlated, then measurement errors become cumulative.)

is the proportion of cases off the main diagonal. For a continuous variable, true change is measured by the stability coefficient, the correlation between the true scores at two time points.

True change occurs when facts and/or people's interpretation of facts change. We will not go into a detailed examination of the nature of attitude change but will specify some factors that will contribute to short-term attitude change. First, attitude change will be greater for questions that are episodic in nature. Episodic questions are those that are closely tied to events that change swiftly and frequently. Prime examples are the presidential popularity indicators and the most important problem questions which can show quite large shifts within a very short interval. All attitude items are presumably influenced to some extent by changing events, but depending on the type of events that are likely to influence a particular attitude, the volatility of the events themselves, and the strength of the association between changes in events and changes in attitudes, items will be more or less episodic. Second, attitude change will be greater when opinions are not crystalized. If an issue is not salient to the public, more people will have weakly held, tenuous positions or be "don't knows." Their positions will tend to change as people waffle back-and-forth, swayed one way or the other by small shifts in events and interpretation. (We can think of their opinions as having relatively little mass. As a result, the opinions can be moved with less effort.) Third, attitude change will be greater when opinions are grouped around the dividing line between sides on the issue. Maximum change will occur when attitudes are distributed as in

Figure A and change will be less likely to occur in either Figures B or C. In Figure A, a maximum number of attitudes can be switched from positive to negative with a given amount of effort since they will have to be moved a minimum distance. In Figures B or C, given the same amount of effort (say moving the dividing line one unit in the positive or negative direction), will switch the position of considerably fewer people.

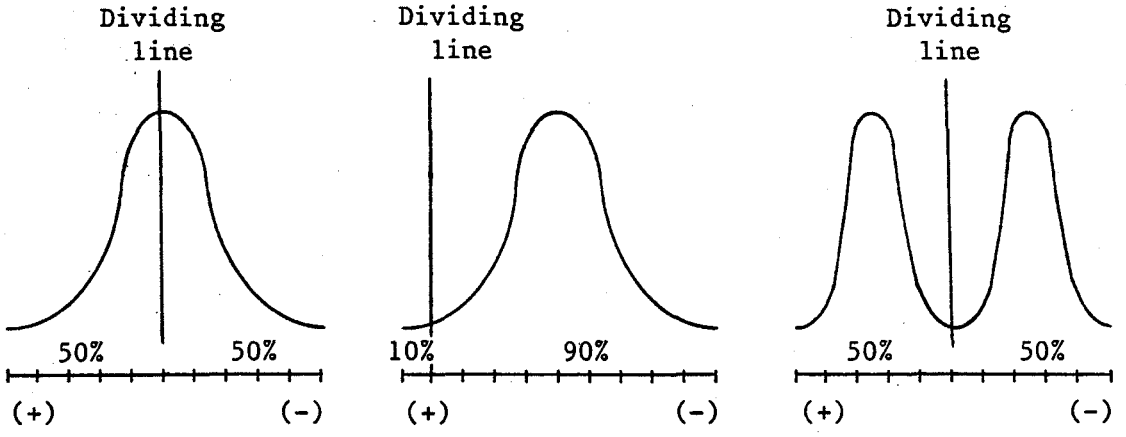

Fig. A Fig. B Fig. C

Finally, attitude change will be greater when there is an inconsistency or imbalance in positions. A normative syllogism (e.g., All virtues are laudable. Kindness is a virtue. Kindness is laudable.) will tend to sustain itself while a contradiction (e.g., All people are created equal. Blacks and whites are people. Blacks are inferior to whites.) will tend to lead to modification of one of the conflicting attitudes. For example, in the contradiction above, if we hold the major and minor premises constant then there would be pressure on the conclusion to change (e.g., Blacks and whites are equal.). Of course, it is not

always the conclusion that will change. The premises can change. In addition, while contradictions are inherently instable, this does not mean that they are always quickly and completely rectified. Contradictions are, however, more susceptible to true change than syllogisms, ceteris paribus.

Using many of the same distinctions we developed above about types of questions, we can describe some of the types of people more susceptible to true change. This would include fence sitters whose opinion is on or near the cutting edge of an issue, apathetics for whom issues lack relevance and saliency, inconsistents who hold contradictory attitudes on a subject, and flighty or manic-depressive people who change moods and opinions easily. In brief, certain issues and certain types of people are more likely to show short-term attitude change than other issues and individuals.

The Entanglement of True Change and Measurement Error

Having elaborated what we mean by measurement error and true change and considered some of the conditions underwhich the two occur we are now brought back to the problem that test/retest or panel data do not distinguish between the two. Taking the simplest case where there is a dichotomy at two time points, there are eight possible occurrences:

		True Change	
		Yes	No
Measurement Error	Time 1	A	B
	Time 2	C	D
	Time 1 & 2	E	F
	Neither	G	H

Now in cases A, C, E, and G, true change occurs, but observed change occurs in B, D, E, and G. Only in cases E and G does the observed change actually represent true change (and in case E, x to y change is observed as y to x change). Cases A and C are observed as being stable because measurement error neutralizes the true change and cases B and D are observed as changing because measurement error changes the observed values although no true change has occurred. Only in cases G and H are true change and stability respectively recorded precisely (although case H in particular would, under most actual circumstances, have the bulk of cases). Two apparent situations result from this configuration. First, many individuals will be wrongly classified as changers or non-changers and since the over-and undercounts of changers need not balance out in any set fashion, even the aggregate level of change will probably be distorted. Clearly, in order to accurately measure true change or measurement error, some individual or aggregate way of decomposing observed change/stability must be developed.

In order to illustrate how measurement error and true change combine together to distort observed change, take the case where there is a dichotomy with 60 percent of the cases in category 1 and 40 percent in category 2. Assume that (1) 10 percent real change occurred and that this was proportional for each category and random, (2) there was 10 percent random measurement error, and (3) the true change and the measurement error were independent of each other. In this situation, the observed data would not show 10 percent change over the test/retest interval but 24.4 percent change. Thus, on the aggregate level, observed change would exceed true change by a factor of 2.44. On the individual level, the

8.2 percent of the cases which were undergoing true change would show change although in 0.1 percent of the cases, the direction would be reversed. In 1.8 percent of the cases, the true change would be canceled out by measurement error and go unobserved. In 16.2 percent of the cases, measurement error would show up as observed change and in 0.9 percent, measurement error would occur but show no change. Finally, for 72.9 percent of the cases there will be no measurement error and no true change. Now, if we combine the 72.9 percent correctly recorded as being constant with the 8.1 percent of the cases where true change was recorded (and in the right direction), we find that in only 81 percent of the cases are things precisely right. Adding in the 1.0 percent where error occurred but there was no misclassification regarding true change and consistency gives only 82 percent of the cases in which a case is accurately classified as changing or unchanging. In sum, on the individual level, 18 percent of the cases are misclassified in regards to true change and, although some of the errors cancel out on the aggregate level, misclassifications are still 14.4 percent.

Test/Retest and Alternative Designs

Before considering how measurement error and true change can be separated in test/retest data, we will briefly consider the general advantages and disadvantages of this research design, compare in general test/retest with some other techniques, and in particular, inquire whether the whole problem of disentangling measurement error and true change can be avoided by using alternative designs that serve the same needs without creating the same problems.

In addition to the problem of confounding measurement error and true change, the test/retest approach is also challenged on several other

grounds (Cureton, 1968; Nunnally, 1975; and Bohrnstedt, 1969). The two major faults are (1) reactivity and (2) non-correspondence between internal consistency and test/retest measures of reliability. Measurement reactivity can occur in several ways. A respondent may remember previous responses and intentionally (for the sake of consistency) repeat them even when true change has occurred or a previous response is known to be erroneous. This can lead to the underestimate of true change and correlated errors between test and retest. Alternatively, a respondent may be socialized by the initial interview experience so he/she is better prepared for the retest (e.g., less anxious). This may reduce measurement error during the retest. Finally, the initial test may even sensitize the respondent to certain issues such that a rethinking of attitudes or an attempt to gain more information on a topic occurs. This may lead to either a different pattern of true change than would have occurred otherwise or to a decrease in measurement error as additional information and thought reduces misunderstandings and some other forms of measurement error. (Of course, other, even opposite, patterns are possible, but these seem the more plausible.) Empirically, in sociological type surveys there are little on the extent of reactivity. The psychometric literature suggests on one hand that some reactivity is common but memory consistency effects decline rapidly with time (Selltiz, 1976; Webb et al., 1966; Sudman and Bradburn, 1974). The literature suggests that for periods of several weeks and longer the memory consistency effect is usually trivial, but the impact and durability of interview socialization and politicalization effects are uncertain. It is quite clear, however, that measurement reactivity can result from the test/retest approach and this factor must be considered when evaluating test/retest reliability.

Another criticism of test/retest reliability points out that there can be little correspondence between test/retest and internal consisitency measures of reliability. If a scale is made up of items that are individually highly consistent over time, then the scale will also have a high level of consistency over time and therefore a high test/retest reliability measure. On the other hand, if the items in the scale are unrelated to each other, then the internal consistency can be very low. For example, an additive scale of year of birth, height, number of siblings, and years of school would probably produce a high test/retest reliability but have little internal consistency. Based on the face absurdity of our proposed scale, one might be predisposed to believe that the internal consistency's evaluation of the scale as unreliable is correct and that the test/retest rating is in error. One might, in turn, extend this argument to state that test/retest reliability is therefore a suspect approach in general. There is, however, an alternative way of looking at this case. One might see internal consistency here as a measure not of reliability, but of construct validity. Under this interpretation, what we have is a reliable, but invalid scale. In fact, if we accept reliability as measuring how many wrong responses to true responses there are, then we have to adopt the second approach as accurately describing the situation.[9]

---

[9]High reliability does not insure validity and this is true of test/retest, parallel forms, and internal consistency. Reliability is how well an instrument measures whatever it is it measures. By "how well" we mean accuarcy or truthfulness of the instrument. A correctly calibrated ruler that accurately measures distance or a national origins question that correctly records country of birth are reliable measures. Validity is how well an instrument measures what it is intended to measure. Thus, using a reliable ruler to measure distance is a valid use or to put it another way, a correctly calibrated ruler is a valid measure of

If we consider turning to alternative psychometric methods of determining measurement error, there is the problem that they are generally not appropriate to typical sociological data. Both parallel forms and internal consistency techniques require a number of multiple indicators or scale for every attribute being measured. (Blalock suggests 10-12 items per scale may be sufficient, but this seems to be pushing the lower bounds pretty hard.) Typical sociological data unfortunately do not, and in many instances cannot, come in this form. It is difficult to imagine how one might determine measurement error for such background variables such as education, religion, or age with a parallel approach and it certainly would be inefficient. For most behaviors and attitudes, it would probably be possible to craft appropriate scales, but this is rarely done. (Only one explicit scale on the General Social Surveys, a ten item vocabulary test, meets Blalock's rule of thumb minimum.)[10] Single item indicators are commonly used (but not universally approved of) in social research and even standard multiple-item scales (e.g.,

---

distance. Using the ruler to measure national origins would be an invalid one. A reliable ruler is not a valid measure of national origins. The relationship between reliability and validity can be shown in the following four-fold table:

		Validity +	Validity −
Reliability	+	A	B
	−	C	D

In cell A, we have the ideal case instrument that is highly reliable and highly valid such as a correctly calibrated ruler being used to measure distance. In cell B, an instrument that is highly reliable is being invalidly used such as an accurate ruler being used to measure wind velocity. In cell C, we have an instrument that is very unreliable and highly valid. This combination is virtually impossible since an instrument that contains a high level of measurement error can not give consistently valid measures. In cell D, we have an instrument that is very unreliable and invalid such as a miscalibrated ruler being used to measure wind velocity.

[10] By contrast, a sample of 20 social-psychological scales from Robinson et al., 1973, had a median length of 30.5 items.

SRC's political cynicism scale, five items, or Treiman's Pro-Integration Sentiments Scale, eight items, frequently reduced to five items) fail to meet the psychometric model. Current sociological practice could be altered to more closely match the psychometric model, but it is not certain that this would be the best approach. Using large scales to measure each construct of interest would clearly lead to better (more accurate and more complete) measurement of the construct and facilitate the calculation of reliability which in turn would improve multivariate analysis of interrelationships. Practically, however, this approach might be undesirable since a single or short-scale indicator of an attribute/attitude might be nearly as good as a long multiple indicator making the extra effort unnecessary. (Ideally, of course, one would be willing to substantially enlarge the scale in order to improve it even minimally.) Also, given fixed resources, increasing the number of items needed to measure a construct would reduce the number of constructs that could be measured. This might lead to trade-offs such as being able either to measure, say, (1) three key concepts very well, (2) six key concepts fairly well, or (3) twelve key concepts minimally.

In addition, adopting the parallel form or internal consistency psychometric methods of measuring reliability does not help with the problem of correlated error. Since these errors are a frequent and occasionally major source of measurement error, the inability of these techniques to unhandle them is a serious limitation. Also, the construction of long multiple-item tests to measure concepts will in some cases actually exacerbate the problem of correlated error by unintentionally creating a response set or other bias. Of course, it should be possible to minimize this possibility, but careful instrument design and empirical testing

will be needed in order to avoid the pitfall. In brief, it appears that other psychometric techniques for measuring reliability are no more panaceas for sociological research than the test/retest approach.[11]

There are, likewise, alternative ways of studying true change: the most prominant being time series cross-sections (several independent surveys over a given time period), cohort studies (a single cross-section in which change is inspected by examining differences between cohorts), and time series cohort studies in which cohorts but not individuals are followed in successive cross-sections. We can not even begin to consider the advantages and disadvantages of studying change through these various approaches. It is clear, however, that only panel analysis permits the examination of individual level changes and thus the examination of net and gross levels of change.[12] If either of these features is important for analytic purposes, another approach will not be fully adequate. In brief, we find that the test/retest design and its accompanying problem of disentangling measurement error and true change cannot be simply avoided by turning to alternative designs that (1) avoid the intermixing, (2) permit the same analytic possibilities, and (3) are not at least as technically deficient as test/retest in other regards.

---

[11] Several scholars have taken a definite position on the merits of the various ways of measuring reliability. Among the anti-test/retest group, George W. Bohrnstedt states, "Because of the problems inherent in the test-retest approach to reliability assessment, many scholars have abandoned measures of stability for what are called measures of equivalence . . ." (p. 86) Similarly, Jum C. Nunnally observes, "Except for certain special instances, there are serious defects in employing the retest method." (p. 335) John P. Robinson, Jerrold G. Rusk, and Kendra B. Head, on the other hand, lament, "It is unfortunate that test-retest measures, which require more effort and sophistication on the part of the scale developer and show lower reliability figures for his efforts, are available for so few instruments in the literature." (p. 16)

[12] Recall surveys can actually do the same, but because of memory effects and other deficiencies, this approach is considered inferior to panel design and is only rarely used in case of attitudes.

Disentangling Measurement Error and True Change

Given that true change and measurement error are unfortunately intertwined in test/retest data and that one wants to disentangle the two, the question becomes how? There are actually several ways in which the two components can be separated, although most of the solutions have definite limitations. First, we will consider whether the problem can be solved by setting one of the components to zero. Second, we will examine solutions where one of the components can not be set to zero. We can further divide these solutions into two major groups: those leading to individual-level decomposition and those leading to aggregate-level decomposition. On the individual level, this means that an individual's responses can be classified as showing true change or measurement error. on the aggregate level, we know the mixture of true change and measurement error for all cases but not for particular cases. Individual approaches include (1) reconciliations, (2) various exploration techniques such as debriefings, interviewer evaluations, and respondent inquiries, (3) verifications, and (4) hybrid methods. The aggregate approaches include (1) three-wave methods, (2) two-wave and multiple indicators, instrumental variables, parallel forms, or experimental treatments, and (3) combinations. In the following sections, we will describe each of these approaches and consider some of their strengths and weaknesses.

One way to separate out true change and measurement error from each other is by making the simplifying assumption that one of the elements is zero. This occurs (usually unwittingly) in substantive panel studies when analysis is carried out without regard for the contamination of measurement error. In studies of reliability, the opposite error occurs when one assumes that no true change occurs. Both of these assumptions

are empirically unwarranted in most situations.[13] In two special cases, however, the assumption of no true change becomes tenable. First, if one is measuring what is defined as an enduring construct, then any short-term variation might be considered as equivalent to measurement error. What you would be doing in this case is having a test/retest correlation measure both reliability and validity. Short-term true change would indicate that the scale was not a valid measure of enduring construct and this, together with random measurement error, would indicate the inadequacy of the scale. Of course, it is not preferable to mix reliability and validity in this fashion and the whole exercise rests on the assumption that there really is a highly stable construct to be measured (and not that we are wrong about its stability). Still, given the difficulty of separately measuring validity and if we accept the assumption of stability, then it might be useful to consider true change and measurement error together in these situations.

The second case in which it makes sense to ignore true change is in the case of unchangable attributes. Certain variables cannot actually change over any test/retest period. These include all fixed characteristics and historical references. Fixed characteristics include sex, race, cohort, and the like. Historical references include items dealing with past events. For example, country of birth, candidate voted for in a given election, father's occupation when respondent was 16 years old, or initial reaction to President Kennedy's assassination. In addition, certain variables are extremely unlikely to change over a short period. Over a one-month period, the following variables would remain stable

---

[13] Some measurement error is probably always present and unless one has strong evidence that it is very trivial, it can not be blissfully ignored.

for virtually everyone: years of schooling, marital status, and number of siblings ever. Of course, one must be careful when going from absolutely unchanging attributes to rarely changing attributes. If one is not very rigorous in the use of this concept, one quickly slips into the common error of ignoring true change because it "must have been trivial." It is probably a generally good rule not to consider current attitudes and behaviors as unchanging no matter how short term the period or enduring the trait. Unchanging attributes will generally consist only of fixed characteristics, historical references, and demographics that are extremely unlikely to change over the period and population being covered. (As a further example, asking a national sample of adults how many siblings they have could change over a month only if their parents had another child. Given the age distribution of their parents and known biological constraints, such occurrences would be very rare, especially during any particular one-month interval. Likewise, asking a school-age sample in September and November how many years of schooling they had completed in full would record extremely little true change, but asking the same population the same question between May and July might record nearly universal true change.) In brief, in a few carefully specified circumstances, it is actually possible to set true change to zero. This, of course, greatly simplifies the interpretation of test/retest data.

The possible application of this advantage and insight that can be drawn from it is evident from a brief inspection of GSS test/retest data (Table 1). Unchanging demographics referring to present conditions had the highest level of consistency, generally over 97 percent giving the same response both times. Next came unchangeable demographics referring to past conditions. Behavioral items apparently come next but there

TABLE 1

MEAN PERCENT CONSISTENT BY QUESTION TYPE[a]

	Unchangeable Demographics			Changeable Demographics	Behaviors	Attitudes	Personal Evaluations
	All	Past	Present				
1972	95.8 (22)	94.6 (12)	97.3 (10)	89.9 (7)	97.2 (2)	84.3[b] (35)	86.1 (8)
1973	92.9 (10)	87.8 (4)	96.3 (6)	91.2 (4)	94.5 (1)	85.8 (27)	84.3 (4)
1974	93.0 (9)	88.5 (5)	98.5 (4)	94.3 (2)	--	82.6 (6)	79.6 (3)
1978	93.5 (2)	93.5 (2)	--	94.3 (1)	85.8 (4)	82.1 (13)	--

[a]The percentage agreeing in this table were calculated on the following basis: (1) all items were dichotomized either into standard categories (e.g., South/Non-South) or, when no standard collapse was obvious, as close to a 50/50 cut as possible; (2) "don't knows" were excluded from analysis; and (3) cases that were asked a question at one time but not at a later time, because of changes on a screener or filter, were excluded from the analysis. This gives figures that in general will be higher (i.e., show more consistency) than if other conventions were used.

[b]Excluding the nine questions that were asked of whites only, the average was 82.6 percent.

are too few observations to be confident about this ranking. The least consistent are attitudes and personal evaluations with consistent responses in about 80 to 85 percent of the cases (see Appendix 2 for details).

Differences in both reliability and stability account for this pattern of consistency. Unchanging demographics about present attributes such as year of birth, educational level, and income refer to concrete, basic, salient attributes with minimal recall required. As a result, reliability for these items is at a maximum level. Since they refer to attributes that cound not possibly change between the test and retest (e.g., year of birth) or were extremely unlikely to have changed (e.g., years of schooling), their stability is essentially perfect. Changeable demographics such as religion, number of earners in the household, occupation, or party identification refer to attributes that are nearly as concrete, important, and standard as the unchangeable demographics. True change is not highly likely over the test/retest interval, but can occur. Thus, they tend to be less consistent than the unchangable demographics because of this possibility for change. Unchangable demographics from the past, on the other hand, have perfect stability, but refer to attributes that are less salient and less concrete and which necessitate recall over a much longer period such as type of community lived in at age 16, father's education, and having ever received government aid. Their lower consistency comes from their lower reliability. Behavioral, attitudinal, and evaluative items rank lower on consistency because their stability and reliability are lower. The items range from high to low on concreteness and saliency, and are susceptible to a greater or lesser amount of true change but are in general probably average lower than demographics on

both stability and reliability. Looking at certain topical groups of attitudes, we can see some of the characteristics that lead to higher or lower consistency:

	1972	1973
Stouffer civil liberties	.756 (9)	.818 (6)
Misanthropy	.793 (3)	--
Spending	--	.863 (11)
Others' attitudes	.869 (4)	.868 (4)
Abortions	.882 (6)	.881 (6)
Race relations	.887 (10)	--
Crime	.922 (3)	--

The two factors that seem to be related to the variation in consistency across these topical areas are the concreteness and saliency of the issue. The Stouffer civil liberty question asks about permitting civil liberties in various situations for various groups. The questions are hypothetical and not addressed to issues that were of headline importance. Misanthropy asks about the basic nature of one's fellow man (rather abstract) and not tip-of-the-tongue salient. The crime and race relation questions, on the other hand, were generally about concrete issues that were highly relevant during the period. In brief, it appears that consistency will be higher on attitude items when the issue addressed is a concrete issue of current topical interest and lower when the issue is more abstract and less salient.

So far, we have seen that there is no clearly superior substitute research design that permits us to avoid altogether the problems of the test/retest design and that setting one of the components to zero, while a very useful technique in selected circumstances, is not applicable in most instances. Since there are no simple solutions, we will consider next some more general and complex methods of separating error and change.

Individual-Level Approaches

Looking at individual-level approaches first, there are four main approaches: (1) reconciliations, (2) exploration techniques, (3) verifications, and (4) hybrid methods.

On the 1972 and 1973 reinterview, reconciliations were obtained for seventeen items. The original responses to these questions were recorded on the reinterview form and if the response on the reinterview disagreed with the original response, the reinterviewer attempted to reconcile the divergent responses. A typical query asked, "Now in the original interview, the interviewer recorded _____ (READ WHAT WAS RECORDED). Now you just told me _____. Could you think about this a moment? Perhaps you could explain why the information is different."

The reconciliations were coded into one of eight reasons:

Code	Response
0	R changed response: admits original response but now a) feels differently b) changed opinion c) has thought about it and wants to give a corrected or different response d) admits to wrong information on original
1	R changed response: admits original response but doesn't know why he/she gave it
2	R changed response: error due to a) misunderstood or misinterpreted original question b) used a wrong frame of reference or a frame of reference not suggested by the question
3	R says guessed on original and guessed on reinterview (didn't have enough information to answer the question or couldn't really choose between alternatives)
4	R changed response: but can give no explanation about why it is different
5	R changed response: gave an answer historically correct (i.e., described R's behavior some time in past but not now correct)

6 Interviewer error:

a) R claims never said that
b) R suggests that interviewer might have misunderstood

7 Interviewer decision:

a) R gave two answers and interviewer recorded one only
b) interviewer made a judgment about R's response

By adding across all items to increase our case base and regrouping the reconciliations in various alternative fashions, the following highlights appear:

1) Most disagreements (.755) are the result of changes by the respondent. Decisions and errors by interviewers, coders, and keypunchers account for about one-quarter of the disagreements (see Table 2 for details).

2) Guessing (either for correct answer or between categories) is credited for .179 of disagreements. Most of the guessing was interviewer or coder guesses between vague or dual responses.

3) Instability appears to account for more disagreement than unreliability. Changes due to altered conditions or opinions account for almost half of disagreements. Changes due to error in understanding, guessing, and decisions and errors by data collection account for about two-fifths of disagreements. A remaining one-tenth is due to changes by respondent that may result from either unreliability or instability, but not from data collection/processing.[14]

4) Most unreliability appears due to data collection rather than respondent (respondent = .138, data collection = .245). However, since all undistinguished unreliability is respondent oriented, the true balance is probably close to equal.

Looking at how items and groups of items varied according to their reasons for disagreement, we find that:

1) Unreliability accounts for most of the disagreements on demographics. Since the demographics refer to attributes that were theoretically fixed (residence at age 16, family standing when growing up, ethnicity, last year's family income, and labor force status at time of original interview) this is hardly surprising. In fact, all reasons indicating instability on these demographics must be considered to be in error.

---

[14] Assessing the share of disagreements due to instability and unreliability is hampered by two factors. First, code "0," counted here as indicating instability also includes some unreliability (reason d and, in part, reason c). Second, the unreliability/instability mixture of the changes in which the respondent changed responses but did not indicate or know why is not known.

TABLE 2

ITEMS BY REASON FOR DISAGREEMENT

Item (Year)	Reconciliation Codes[a]								Total
	0	1	2	3	4	5	6	7	
RES16 (72)	3	0	2	1	0	0	4	1	11
INCOM16 (72)	2	1	1	2	2	0	3	1	12
WRKSTAT (72)	3	0	1	0	0	1	0	2	7
WRKSTAT (73)[b]	3	2	3	0	0	29	6	19	62
COURTS (72)	5	0	0	0	0	1	2	0	8
HEALTH (72)	5	0	1	0	4	2	1	0	13
SATJOB (72)	16	4	1	0	1	12	2	0	36
FINRELAT (72)	5	1	1	1	0	1	0	1	10
HAPPY (72)	2	2	1	1	0	5	1	1	13
NEWS (72)	2	0	0	0	0	1	0	0	3
ETHNIC[c] (72)	7	0	1	2	1	1	6	6	24
CHLDIDEL (72)	5	1	2	1	2	0	1	5	17
ATTEND (72)	4	1	1	0	1	4	3	1	15
ATTEND (73)	6	0	8	2	6	9	4	3	38
USINTL (73)	13	1	3	4	3	0	2	1	27
INCOME (72)	10	0	5	1	3	1	1	1	22
Total	91	13	31	15	23	67	36	42	318

[a]See previous discussion for definition of codes.

[b]In 1972, WRKSTAT referred to what the respondent was doing the week of the original interview. In 1973, it referred to the labor force status "last week" or, in other words, a week before the re-interview or about five to six weeks after the original interview.

[c]Questions also used to code variable ETHNUM. This variable is not analyzed here.

2) Instability is greatest for personal evaluations (happiness, job satisfaction, health, financial position) and followed by attitudes and behaviors. It is also high for demographics such as labor force status in 1973 where the time reference shifts.

3) Guessing is high for items with a large number of closely related responses (e.g., ideal number of children) and/or relatively complex coding rules (ethnicity and labor force status).

In sum, the reconciliation data confirm that both unreliability and instability are major causes of disagreement on test/retest comparisons. Unreliability comes both from guesses and errors by respondent due to misunderstanding or uncertainty and from guesses and other errors in the data collection process. Instability comes from actual changes in conditions and attitudes. The relative contribution of unreliability and instability varies according to the type and form of question asked. For demographics, both by definition and in practice, disagreements arise from unreliability. In addition, unreliability in the form of guessing comes from items with a large number of closely related categories or complex coding rules. Attitudes, behaviors, and especially personal evaluations have more disagreements from instability. Instability also accounts for many disagreements when the time reference changes.

A particular advantage of the reconciliation approach is that it permits the identification of true change and measurement error for individuals. Rather than working from the aggregate, one can identify particular individuals who were stable and/or reliable and use this individual-level information to adjust the data (e.g., measuring true change by excluding observed changes, or inconsistencies, that were due to measurement error), and to study the particular reason for measurement error, and to examine the attributes of individuals showing either true change or measurement error.

Reconciliation will, of course, not pick up any constant measurement error. This includes both uncorrelated error that by chance produces two wrong responses and correlated error occurring in both administrations such as a social desirability effect or response set. Even among the discrepancies that trigger the reconciliation, the procedure will not always determine the true reason for the inconsistency because there will be measurement error (both correlated and uncorrelated) in the reconciliation process.[15] Despite this added error, it should be possible to determine the true change for inconsistency and accurately disaggregate most inconsistencies into measurement error and stability components. There will, thus, be a large net gain in precision although consistent measurement error between test and retest and measurement error during reconciliation will prevent complete accuracy.

Another direct inquiry method of assessing reliability and thereby helping to separate the chaff of measurement error from the grain of true change is the debriefing technique. Debriefing covers a multitude of types that all involve asking the respondent about his/her understanding of the question or elements of the question. This may consist of asking the respondent to define terms used in the question, to elaborate upon answers such as asking why a particular response was given, to verify asserted knowledge of a fact by answering specifics, or to check over initial answers to make sure that the correct response was registered. In general, the debriefing technique is used to determine how respondents are interpreting and comprehending an item and to locate cases where measurement error is occurring because of misinterpretation or other

---

[15] As when disagreements on unchanging demographics were credited to instability.

problems. Most debriefing techniques are distinct from the test/retest approach although they obviously complement its attempts to calculate measurement error. The check-over technique, however, has certain similarities to an instantaneous test/retest design. The basic similarity is that the respondent goes over the same questions two times in both approaches. The differences are probably greater than the similarities, however. In the test/retest approach the questions are asked a second time and minimal reference is made to the initial asking of the questions. One hopes that memory and other lag effects are nil so that the retest is essentially an independent measurement of the attitude/attributes in question. In the debriefing approach, the respondent is not only given the questions again, but the answers as well. The respondent looks over the recorded response and checks to see if these are the answers he/she wants. If a response is found to be wrong, the respondent is asked to supply the correct answer (without the initial response being destroyed) and asked why the change was made (similar to reconciliation). This does not directly help to separate out measurement error and true change, but can help the general assessment of measurement error and, if used in conjunction with a test/retest design, would help to specify cases where inconsistencies were due to measurement error rather than true change.

Interviewer evaluations consist of having the interviewer rate the respondent on certain attributes that are related to measurement error. The interviewer may rate the respondents' understanding/comprehension, their cooperation, their frankness/truthfulness, or some related attributes. For example, on the GSS, interviewers are asked, "In general, what was the respondent's attitude toward the interview?

Friendly and interested, Cooperative but not particularly interested, Impatient and restless, or Hostile," and "Was respondent's understanding of the questions . . . Good? Fair? Poor?" Usually these evaluations are done at the end of the interview and are global in nature, but they can be used to refer to specific sections or items. Another way of having interviewers evaluate specific parts is illustrated by the following question from the 1976 American National Election Study of the Survey Research Center, "Were there any particular parts of the interview for which you doubted R's sincerity? If so, name them by section or question numbers." The chief drawback of this approach is that there is bound to be a lot of error in interviewers' evaluation of such matters as comprehension, veracity, and cooperation. In addition, the application of post-interview general assessments to particular items or scales is obviously dubious. Still, the technique may well help to identify problem (unreliable) individuals and questions and, if judiciously applied, could help to disclose measurement error.

The final method is the respondent inquiry approach. It might be considered as part of the debriefing technique, but we have decided that it has certain distinctive features that make it worthy of separate mention. In this approach, respondents are asked the same type of questions that interviewers rate respondents on in the interviewer evaluation approach. Some examples from NORC 5025 include:

> Overall, would you say you enjoyed the interview very much, somewhat, or not at all?
>
> Were any of the questions unclear of hard to understand? Which ones?
>
> How about the questions on voting--do you think they would make most people very uneasy, somewhat uneasy, or not at all uneasy?
>
> Do you think the voting questions would annoy most people--very much, somewhat, or not at all?

A typical problem with this approach is that one has to gingerly approach the topics of comprehension and (especially) veracity. It is doubtful that one can boldly ask, "Did you lie about this question?" so one tries to indirectly measure propensity to lie by measuring the degree of threat (uneasiness and annoyance). The problems are (1) whether you accept the respondent's evaluation to be any more truthful than his actual substantive responses to questions (for example, if he didn't understand the question, will he admit his ignorance and, if he lied, will he admit his falsehood); (2) if such indirection is used, whether one can determine if false answers were really given or does the indirection make it impossible to determine this.

The debriefing, interviewer evaluation, and respondent inquiry methods can be used in several ways to help assess test/retest results. If the techniques are used at both times (test and retest), then certain patterns of measurement error can be discerned and either the data can be adjusted to reflect true scores or, where this is not possible, the errant cases can be segregated for separate analysis. If the information is available for only one time point, it is difficult to determine the pattern of measurement error unless certain assumptions about the likely replication of measurement error are made. Since such a procedure would be rather questionable, it would be preferable to use the known data on error to separate out the test/retest cases for comparative analysis.

One of the particular advantages of the debriefing, interviewer evaluation, and respondent inquiry methods is they can detect constant error. Reconciliation, as we noted above, doesn't come into operation unless an inconsistency occurs. This excludes any possible consideration of consistent error. Since these techniques can (at least sometimes)

ferret out constant error, they can greatly complement the reconciliation approach and refine the analysis of measurement error and true change.

A third way of separating measurement error and true change is by verification. Responses are checked against objective standards which can confirm or refute the veracity of the response. For example, one might check a respondent's claim that he is registered to vote against the voter lists (Traugott and Katosh, 1979; Bradburn and Sudman, 1979) or membership in particular voluntary association against the membership rolls. By verifying the true condition at both test and retest periods, one can obviously determine whether any measurement error or true change occurred. The drawbacks of verification are (1) it is applicable only for attributes that are subject to checks against objective standards (this excludes all attitudes, most behaviors, and some demographics); (2) in order to accept the objective standard (voter list, membership roll, employer's report, etc.) as the arbitrator of truth, one must assume that it is inerrant (or at very least more accurate than the respondent); and (3) verification is costly. In those cases where objective standards are available for cross-checking, it can be a powerful technique for uncovering measurement error, especially certain types of nonrandom error. In turn, this information can be used to separate out measurement error and true change in test/retest data.

Finally, one can devise a wide variety of hybrid methods of separating measurement error and true change by combining together two or more of the individual-level techniques. Without considering the numerous possible hybrid methods and their particular pluses and minuses, we will simply note some general features of the hybrid approach. First, using two or more approaches rather than one will almost always increase the

proportion of errors that can be identified. Second, since reconciliations can not detect constant error nor the situation when measurement error cancels out true change to leave the appearance of consistency, it would be beneficial to use one of the alternative methods that can detect these situations along with reconciliation. Third, since increasing the amount of resources that are devoted to finding measurement error will usually mean fewer resources available for measuring other elements, one must consider the trade-offs that are involved. In sum, there are various techniques by which measurement error and true change can be separated on the individual level. Unfortunately, no approach is complete, cheap, and simple. Only with considerable and careful effort can measurement error and true change be sifted apart and even then the refinement will never be total.

Statistical Approaches: Overview

The statistical separation of measurement error from true change is simple if the amount of error can be estimated by some method, such as parallel forms, which does not involve a lapse of time with the consequent possibility of true change. Once the measurement reliability is known, stability can be determined from a simple two-wave panel. The usefulness of parallel forms, however, seems limited to situations such as in psychological testing, where one devotes a great deal of effort to measuring a small number of fairly abstract concepts. The reliability of sociological measures must usually be examined by repeated measurements, so that both measurement error and true change contribute to observed inconsistency. One needs a model which incorporates error and change, and which permits their separation.

Over a decade ago, Coleman (1968, pp. 453-454) pointed out that under certain circumstances one could distinguish true change from random measurement error in panel data with measurements at three time points. Heise (1969) showed how this insight could be translated into the well-known path-analytic representation of a linear system. The ensuing decade has seen numerous attempts to generalize or improve Heise's model. We shall not attempt to review all of this literature (see Wheaton, Müthen, Alwin, and Summers, 1977,[16] for a careful discussion of the main developments), but shall concentrate on certain problems which have not been emphasized in the methodological literature. None of these panel techniques may be safely regarded as a generally applicable, robust way of summarily describing measurement error and change. The models, particularly in their more sophisticated versions, are steps toward an adequate treatment of measurement error, but are not adequate, in any real sense, for the solution of present problems.

We shall begin by presenting the basic model for assessing stability and reliability in three-wave data, and its path-analytic interpretations as given by Heise and by Wiley & Wiley (1970). We will then consider in more detail some of the suggestions for further developing the model, in particular by using multiple indicators and by more elaborate theoretical specification. We intend by this discussion to show that these refinements are inappropriate to our present purposes, and also that they have more

---

[16] This article is the most sophisticated discussion aimed at sociologists in the present methodological literature on reliability and stability. Because of the power and generality of the techniques developed by Wheaton et al., we will frequently focus our critical comments on their article. We hope to highlight some of the issues which will arise in empirical application of the techniques, but which are either neglected or taken as implicit in the methodological literature.

general drawbacks not often discussed. Next, we will briefly describe a model for discrete variables developed by Henry (1973) in analogy to the continuous-variable path model. We will conclude with a discussion of problems arising from correlated measurement error. Some analysis of the GSS data will be produced in argument that error correlations are neither infrequent nor trivial. Their existence has serious implications, we believe, for studies of sociological measurement error in general.

A preliminary word is in order about problems particularly important to sociological measurement theory, and particularly the problem of how one knows what is being measured. In the psychometric literature on testing, where sophisticated treatment of measurement error entered the social sciences, this problem was not urgent. Difficult problems arise in determining precisely what sort of ability one was measuring, to be sure; but after all test performance is fairly direct evidence of some sort of ability. Furthermore, the tests can be and are refined by experimental studies using alternate forms and various evalutions of criterion validity. In sociological or social-psychological measurement the problems of validity are often more severe, and the solutions more difficult. Concepts--even highly abstract concepts--employed in sociological theory are grounded in everyday experience: their connections to the objective world are strong relative to those of psychological concepts. These connections limit one's freedom in choosing an operational definition for use in research. Operational definitions which wrench concepts away from the context wherein they originated result in models that are formal and empty. In a theory with great predictive power, one can forgive the lack of familiar connotations, as the very power of the theory justifies, post hoc, the use of concepts we do not fully

understand. Sociological theory is not like this. Its predictive power is limited, but it maintains its interest because of its close ties to familiar and important questions. One cannot with impunity ignore the problem of a measure's validity, or, to look at things from the other side, the "semantic problem" of determining what an abstract variable means. Abstraction is a necessary part of building any scientific theory; but in sociology one should take care that the abstraction does not become an excuse to treat the theory as empty formalism.

Extended discussion of such matters is seldom fruitful, and the temptation is strong to get on with empirical work as well as one can, allowing for a certain unavoidable vagueness of concepts. This practical attitude is often fully justified. Most sociological variables have high "face validity," that is, their meaning is reasonably clear. Furthermore, the source of the data is usually known--perhaps a survey question--so that the validity of its interpretation is available for challenge. However, a growing amount of methodological literature treats variables which are not directly observed, but are rather analyzed by means of their postulated relationship to imperfect indicators or measures. The interposition of these unobserved variables, or "constructs," splits the question of validity into two: how well the construct represents the theoretical concept for which it stands, and how well the measure represents the construct. The latter question (that of "construct validity") is amenable to statistical treatment; but in our terms it is more a matter of reliability than of validity, since it has nothing to do with the question of what is being measured. In most formulations construct validity is simply the square root of reliability. The remaining part of the

original question, the extent to which the concept of interest is captured by the unobserved variable or construct, is rarely discussed in the methodological literature, since it is not a statistical question. Neglect of this, the real question of validity, is particularly unfortunate in models which suppose the construct to have more than one indicator.

Specification of which indicator or indicators are to be taken as measures of an unobserved variable, and of the precise way in which the indicator(s) depend upon the unobserved variable, constitutes the measurement model for that variable. One specifies a measurement model so that the problems of imperfect measurement can be formulated precisely. However, the gain in precision may be accompanied by a loss of contact with the real world. It is obvious that empirical statements about unobserved variables depend not only upon the actual data but upon the way in which one connects the variables to the data. Such a connection in effect establishes an operational definition: to postulate a connection is to define a construct, and to modify the connection of the variable with any portion of the data, even in roundabout ways, is to redefine the construct. We make these points only because, unless great care is taken, the powerful and flexible techniques available for analyzing unobserved variables permit one to say a great deal about nothing in particular. Despite all the talk about relaxing the assumptions and generalizing the models, there has been a reluctance to admit that "empirical" statements about unobserved variables can never be made routinely, but will always rest on precise operational definitions, which in fact are usually made for the sake of convenience, and are frequently excused as conventional. The problem is particularly dangerous because the technical sophistication lulls one into believing that measurement problems have been solved.

Statistical Approaches: The Basic Three-wave Model

The most direct way to distinguish measurement unreliability from true change is to use some technique, such as parallel forms in psychometric testing, which estimates reliability independently of change. If reliability is known, one can easily estimate the amount of true change between waves of a panel study. Where no such technique is available, as for typical sociological variables, one must usually resort to some form of the strategy suggested by Coleman, which depends upon a panel design with at least three measurements of the same variable.

The basic three-wave panel model is that of Figure D. Here T_1, T_2, and T_3 represent the true variable of interest at three time points; M_1, M_2, and M_3 represent measurements of it at those times; u_2 and u_3 are disturbance terms responsible for change in the true variable; and e_1, e_2, and e_3 are independent stochastic terms representing measurement error. The measurement model is simple: M_i is expected to be a linear

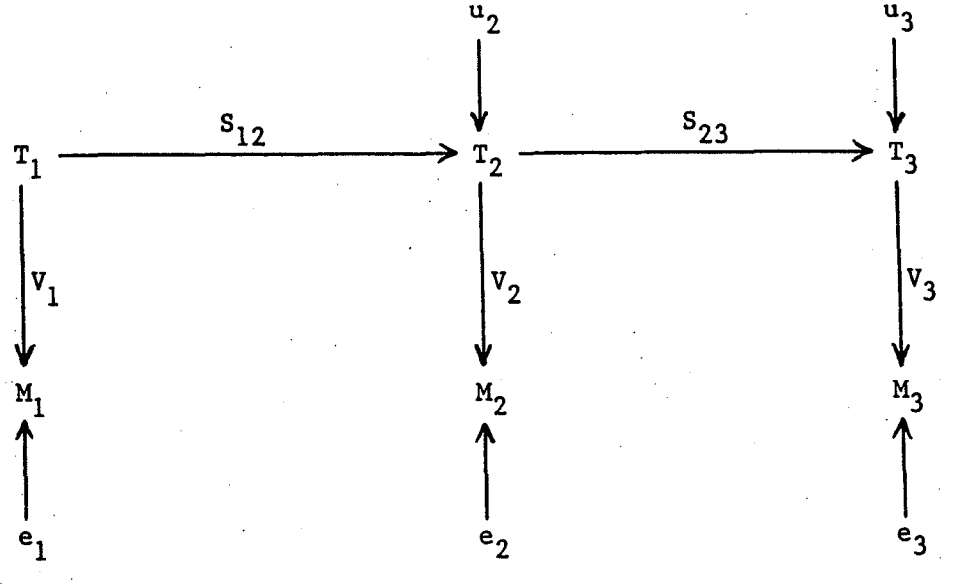

Fig. D. The basic three-wave model.

function of T_i, except for the stochastic error. The specification in Figure D assumes, among other things, that there are no indirect or spurious paths relating the values of the true variable at different times, and that correlations among the measured variables M_i are due entirely to the path connecting them by way of the true variables T_i.

Heise (1969) gave the solution for the standardized path model of Figure D. It is straightforward and we need not reproduce it here, but the basic principle behind the three-wave solution is simple and should be understood. If nothing is changing and measurement error is constant, then the relationship between any two measurements should be the same except for random variation in the error (there is no sampling error in comparing the true scores since this is a panel). Hence the three-wave techniques are not needed when no true change is occurring. On the other hand, if change takes place in some reasonably simple manner (such as all of these techniques assume), then the change between measurements 1 and 3 will be greater than that between 1 and 2 or between 2 and 3. Even with random measurement error confounding the change, one can assess the magnitude of true change between, say, 1 and 2 by comparing the observed (1,3) relationship with the observed (2,3) relationship. Each is attenuated (twice) by measurement error, but the former is weakened also by change between measurements 1 and 2, while the latter is not. Measurement error itself is estimated by looking at discrepancies across the whole interval (including whatever change took place from time 1 to time 3), first with the (1,3) relationship which includes measurement error twice, and secondly with the (1,2) relationship combined with the (2,3) relationship, including measurement error four times. (Error at time 2 weakens <u>both</u> the (1,2) and the (2,3) relationships, so that it

affects the composite relationship twice.) Thus, in the Heise model, r_{13} should equal the product of r_{12} and r_{23}, were it not for the fact that each of the latter two correlations is attenuated by measurement error at time two. The ratio of this product to r_{13} is therefore the square of the path coefficient between the true variable and its measure at time two; this squared coefficient is conventionally called the reliability. Despite great differences in the details, this general strategy is the foundation of all the panel methods for distinguishing between true change, which is independent of measurement, and error which arises from measurement.

For the sake of clarity, we should remark that true change in the form of rapid fluctuations (so rapid that the net effect is practically to randomize the true-score distribution over the shortest test-retest interval) will be interpreted by these techniques as measurement error. The criterion used to distinguish change from error, in this formulation, is that true change persists over time intervals of the order of magnitude of the test-retest interval. Imagine, if you can, a perfect measure of something extremely unstable, such as the number of persons within a hundred feet of the subject. Test-retest correlations could well be positive (some people live in cities), but would probably be as great over a period of two weeks as over one week. One would conclude, erroneously, that the variable was stable but unreliably measured. The error results from the properties of the variable being unlike those assumed in the model. If one chose the conceptual variable "mean number of persons within a hundred feet," and took as its measure the actual number at the time of measurement, the properties of the new conceptual variable would come into line with those of the model, and it would be correct

to conclude that one had a fairly stable but unreliably measured variable. If the correlations happened to decrease over time (as with school children examined in March, May, and July) the decrease would now be correctly interpreted as true change in the mean number of persons in their proximity. This example is farfetched, but the general point is important. The assumptions made about measurement error and true change determine what sort of change, and what sort of measurement error, one can detect.

Wiley & Wiley (1970) have argued that Heise's assumption of constant reliability should be replaced by one of constant error variance. (Some sort of assumption regarding constant error is clearly essential to the above procedure, which relies on counting the number of times error enters into various comparisons.) They were able to do this by using the covariance matrix instead of the correlation matrix, and identifying Figure D as an unstandardized model. The Heise assumption implied, by contrast, that in the event the true variance changed, error variance would change proportionally to it. Although it demands more information, the Wiley & Wiley assumption is not strictly weaker than Heise's, only different. (Ironically, the data they used for their example, on reported earnings, provide an instance where it is not unreasonable that error variance might change proportionally to true variance.) Error variance is quite clearly a property of both the measuring instrument and the population to which it is administered, and not, as they claim, of the former only(Wiley & Wiley, 1970, p. 112). The issue is thus less clear-cut than they imply.[17] By leaving the data unstandardized, the Wiley

---

[17] The Wileys' most forceful criticism of the Heise assumption of constant reliability is simply wrong, as Heise pointed out in his comment (Heise, 1970). This correction should have been included or incorporated when the Heise and Wiley & Wiley articles were reprinted in Blalock (1971).

& Wiley model gains whatever additional information is in the observed variances, so that it will give different results from the Heise model if and only if these change. It will give better results if changes in the observed variances are primarily due to changes in the true, and not the error variance; and we must remember that the latter, but not the former, is liable to change from sampling error.

We suspect that the Wileys' assumptions will be more realistic than Heise's in most applications involving substantial time intervals. Our data span only a few months; and we believe (having looked at the results of both models) that over such a short period changes in the variance are better attributed to fluctuations in measurement error than to changes in the variance of the true scores. Rather than estimate three separate reliability coefficients from this noise, we prefer to suppress it with the simpler Heise algorithm.

We may observe, in passing, that it is mathematically possible to substitute an "instrumental variable" for the first of the three measurements, and thereby obtain reliability and stability estimates from a two-wave panel. Relibility, as always, would be obtained at the intermediate point, which in this case would be the first measurement of the variable one is interested in. Stability could be estimated exactly as before. This possibility seems of dubious value. An instrumental variable to be used with some other variable in a two-wave panel would have to be related to that other variable at the first time, but not at the second time, except by way of its effect at the first time. Since we are talking about the very same variable at both times, this is quite a stringent requirement. We shall not discuss it further.

More Elaborate Models

A number of extensions have been proposed to the basic model. Generally speaking, the suggestions have been to collect data in more than three waves; to use several indicators in measuring each conceptual variable; and to generalize the notions of stability and reliability by modeling several conceptual variables simultaneously. In the following section we shall explain why we adopt none of these extensions. Our objections are not to the techniques themselves, but to their use in answering empirical questions with real data. We feel that many of the models, particularly those based on multiple indicators for the conceptual variables, are more properly considered as tools for exploratory theory construction than for empirical discussion of the real world. Their use is not appropriate unless one is willing to assume the responsibility of imposing considerable structure on the data.

To begin with, it has been remarked that certain assumptions can be relaxed when measurements are available from more than three occasions (Werts, Joreskog, and Linn, 1971; Henry, 1973). The gains are relatively subtle: for instance, one can estimate error variances at times other than the first and last, instead of assuming constancy. Furthermore, the additional data enable one to see how well the assumptions are met. Since our panel data were collected in only three waves, we shall not discuss these variants further.

Secondly, multiple-indicator models (that is, models including more than one indicator or measurement of the conceptual variable(s) of interest) have been advanced with enthusiasm (Blalock, 1970; Wertz, Linn, and Joreskog, 1971; Wheaton et al., 1977). The introduction of

additional measured variables into the model greatly increases the amount of information in the covariance matrix, which may be variously employed for estimating additional parameters or for checking the assumptions of an overidentified model by making redundant calculations. In particular, Wheaton et al. have exploited this extra information to incorporate into their models error terms correlated among themselves, or with the true variables; and also to address more carefully the issue of theoretical specification, which we shall take up later. On a simpler level, multiple indicators permit one to separate reliability and stability with only two waves of panel data.

The simplest reason for our avoidance of multiple-indicator models is the focus in this paper on characteristics of the measures specifically included in the General Social Surveys. Insofar as possible we wish to discuss these measures without imposing any particular theoretical structure, and multiple-indicator models require very precise knowledge of the relation between theory and measures. More generally, we believe that severe problems of estimation and interpretation arise when multiple-indicator models are used in measurement theory, where, if anywhere, it is important to have a clear understanding of what one is measuring.

The power gained by using multiple indicators is based upon one's willingness to correlate the extra indicators with all of the other measures in the model, while specifying that these correlations arise from a particularly simple structure. Frequently one assumes that one indicator is related to others solely by virtue of its dependence upon a single conceptual variable in the model. Here the addition of a new indicator to a model with n other measured variables provides n new correlations, but requires only one more parameter estimate.

The gain in information from the new indicator outweighs the increased complexity of the model only because one has imposed such a simple structure on the connection between the indicator and the rest of the model.

Estimation of such models from sample data poses a dilemma. Unless the various indicators of a "construct" are very highly correlated, one or both will include a substantial proportion of error variance, that is, variance arising somewhere other than the construct being measured. This would mean, first, that one is probably not measuring the construct very well, and more importantly, that the error variance may well be related to something else in the model, which violates the very assumption that makes multiple indicators so helpful. For these (obvious) reasons, it is desirable to have very reliable indicators.

On the other hand, the extra information gained by throwing another variable into the sample correlation or covariance matrix decreases sharply if the new variable is highly correlated with another variable in the matrix, such as the other indicators of the same construct. This is essentially the multicollinearity problem in another guise, and it arises from the presence of sampling error in the correlations. If two indicators of a construct are highly correlated, all of the information gained from the second will be extracted from the slight differences between the correlations of the two with other variables. These small differences, as observed, are disproportionately affected by sampling variation. The resulting instability of parameter estimates can spread through the model in complex ways.

Multicollinearity is simply an estimation problem, although it is one that is quite likely to arise from multiple-indicator measurement

models. More serious, in our opinion, is the question of how one knows what is meant by the unobserved variables or "constructs." In one of the early articles on multiple indicators, Costner (1969, p. 254) cautioned that these do not solve the "semantic problem" of attributing meaning, and that the coefficients of such measurement models "have no bearing whatsoever on the appropriateness, in terms of conventional meanings, of the terms that are attached to the abstract variables." Subsequent methodological literature has scarcely paid lip service to Costner's warning. Analysts doing empirical work must take full responsiblity for any interpretation they bring to constructs based on the powerful factor-analytic techniques used in multiple-indicator models. As to the methodologists, people sophisticated enough to develop multiple-indicator measurement models surely are aware of the all-too-common fallacy of concluding that a certain variable is the best measure of a concept because it has the highest principal-factor loading out of a group of variables which seem to be measures of the concept. Yet this, essentially, is what one does in applying multiple indicators to questions of measurement reliability. The reply that one is measuring "constructs," not concepts, is weak: because postulating constructs is just an ad hoc way of dodging the semantic problems; and because one always ignores the unpleasant fact that the meaning of a construct changes with virtually any respecification of one's model. Constructs of this kind, however convenient their analysis, are simply not interesting things to talk about. Their connection with the concrete social world is gone, and one could summon interest in them only if they proved to have great predictive power. Such proof has not been produced.

Thinking through the implications of constructing, or imposing, a multiple-indicator model upon its empirical base drives home the point

that specification of such a model, in all its detail, amounts to a definition of the construct itself. The successive models estimated by Wheaton et al. (1977) illustrate the ambiguity which creeps in when one acknowledges the definitional character of multiple-indicator techniques. They analyzed a construct called alienation, which they measured with two scales called anomia and powerlessness. Estimating first a three-wave model with uncorrelated errors, they found that, as a measure of alienation, anomia was slightly more reliable than powerlessness. (This is a matter of reliability, not of validity as we use the term. The question of validity, of how well the construct to which they assigned the name "alienation" corresponds to any of the uses sociologists have made of that term, can scarcely be formulated, much less answered. Each revision of the model, and they discuss eight distinct multiple-indicator models in the course of the article, redefines the construct and hence the question.) This statement about the reliabilities of the two measures meant, approximately, that the proportion of the temporally stable variance of anomia which was shared by powerlessness was greater than the proportion of the temporally stable variance of powerlessness which was shared by anomia.[18] They continued by estimating correlations between error terms for the three anomia measurements, and also for the three powerlessness measurements. The former they found to be fairly large, and the latter insignificantly different from zero by their statistical criteria. (This simply meant that anomia

---

[18] Our nominalist interpretations of these models are not intended to be precise statements of what the models do with the observed data. Precision would require much more complicated statements, highly specific to the model considered. A relatively simple complication, for instance, is the presence of another construct, SES, with two indicators, which requires some such phrase as, "controlling for the stable variance common to Duncan's SEI index and education."

had some temporally stable variance not shared with powerlessness.) The effect of introducing these correlated errors for the anomia scale was to shift some of the covariance between anomia measurements away from the path via the construct alienation, where it implied that anomia measured alienation, and into the new path via the correlated errors, where it did not imply any such thing. The "reliability" of anomia as a measure of alienation was thus reduced, so that, as it happened, Wheaton et al. (p. 127) concluded that, "In fact, powerlessness is the more reliable measure of alienation . . . " This conclusion about the relative reliabilities of the two scales sounds like a statement about the measurement of the familiar, if elusive, concept "alienation"; some people would even think that it refers somehow to concepts denoted "anomia" and "powerlessness." On the contrary, the statement derives in its entirety from the implicit definition of the word "alienation" as the stable variance shared between two scales named "anomia" and "powerlessness" in the context of a rather particular model. If a third measure of alienation were available, the conclusion might change again, and the meaning would certainly change; for then one would be talking about a different alienation, namely the stable variance common to all three measures.[19] Evidently the conclusions that one draws from a multiple-indicator model depend very heavily upon the way theoretical constructs are connected with measured variables.

---

[19] Wheaton et al. (p. 108) note that under certain circumstances their estimation technique does not need the assumption that the paths from construct to indicator are constant over time, an assumption which "amounts to specifying that each of the measures . . . are the same measures (sic) across time." Indeed without this restriction, one could use entirely different measures at the different times. This reckless possibility--which they do not advocate--illustrates the fact that one can "estimate" almost anything by throwing indicators into the covariance matrix and drawing a model.

In the present work, we are concerned with the measured variables themselves, and we have no motivation for specifying theoretical constructs precisely and with conviction. We shall therefore refrain from using the multiple-indicator methods, whose interpretation depends upon such delicate specifications.

A third refinement which we think inappropriate is the specification of relationships between different conceptual variables, as discussed by Heise (1969) and in greater detail by Wheaton et al. (1977). Issues of theoretical specification are thought relevant to measurement theory chiefly becuase stability estimates for a conceptual variable will be biased by the existence of other variables causally prior to it.[20] Thus, in Figure E, some variable S (here assumed perfectly stable for simplicity) influences the values of a variable T at all three times when T is measured by the indicator M. If we do not explicitly include S in our model, we will attribute all of the covariance between, say, M_1 and M_2 to the route directly through T_1, T_2, and the "stability" path connecting them; whereas

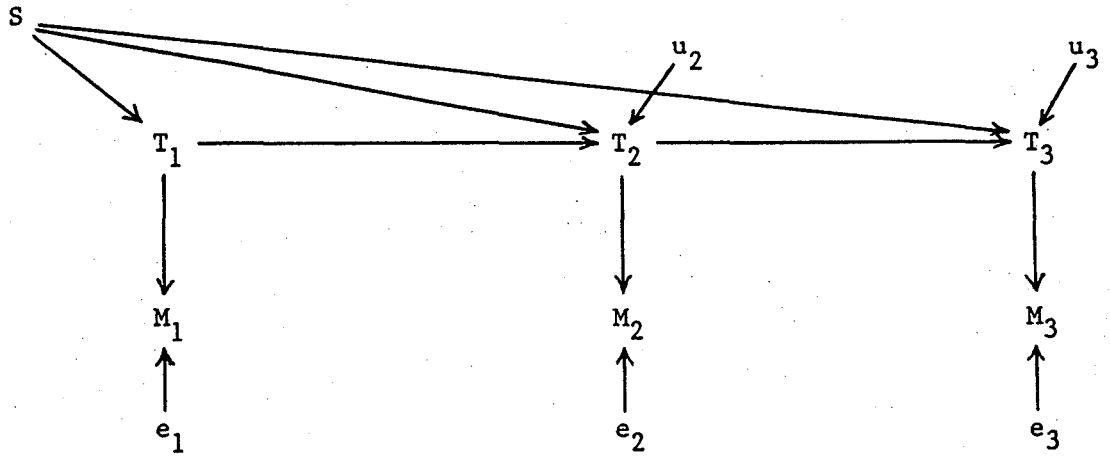

Fig. E. Effect of prior variables.

[20] If this sounds like a very sweeping statement, it is. Wheaton et al. discuss the question at some length, but do not remark on the extremely broad implications of their point of view. We shall note some of these here.

some of this covariance properly belongs to the route by way of T_1, S, and T_2, which bypasses the stability path T_1T_2. Wheaton et al. consider consequences of such prior variables in some detail.

Their conclusion (p. 91) is that if one wishes to obtain "estimates of the 'true' stability" of some variable, one must "first explore the possibility that certain concomitant variables may be causally related to the single variable" in question. This is certainly a good thing to think about, but hardly something to estimate. In scientific research the possibility to which they allude is normally taken to be a certainty, of almost axiomatic character: things have causes. The attempt to assess the stability of a variable independently of all its causes will be futile.[21] In fact, it is not by any means self-evident why one would want to evaluate the stability of a variable in itself, independent of all its causes; nor is it clear what such a pure stability would mean.

We certainly concede that it may be useful, in some instances, to control for the stability or (more often) instability coming from one or two prior variables, provided that one does not speak of completely-specified systems or similar illusions, and that one has a clear conception of what he means by the purified stability parameters. In particular, the example in the Wheaton et al. paper seems a sensible application of their techniques, and an argument for the occasional usefulness of these. As we have remarked above, however, we do not have occasion here for distinguishing a variable's stability from its stability independent

---

[21] Wheaton et al. concluded that their stability estimates were "bottoming out" when, after some exploration, they incorporated into their model the prior variable SES, measured by Duncan's SEI and by education. They presented no evidence for such a conclusion, nor is it clear what evidence could be presented. The number of possible variables prior to alienation must be very great indeed, and no one will have measures of all of them. When to include prior variables, and when to stop including them, are not methodological but substantive questions.

of some specific cause. We are considering variables as they are measured on the GSS, and we prefer to speak of their stability regardless of whether it is inherent or transmitted.

Categorical Data

Thus far we have discussed techniques for the analysis of variables which take on essentially continuous values. For such variables one naturally defines the measurement error as the difference between true and measured scores. If scores are not such that these differences may be meaningfully compared, the path-analytic techniques do not apply. We shall now consider variables which have no such scale, taking on discrete values. We shall speak only of dichotomies. The matrix algebra behind Henry's technique generalizes easily, but would require a much larger case base for polytomies, in order to estimate the necessary conditional probabilities. A dichotomous measurement, then, yields either the right answer or the wrong answer; although one does not know which answer is right for a particular case, one must presume that there is a right answer (a "latent state") if one is to define measurement error. Reliability, defined as the proportion of measured variance which would be "explained" by the true variable, is not a concept relevant to dichotomies. One is instead interested in the proportion of measurements which are wrong, or, on the individual level, the probability that a person will give the wrong response. In this context a reliable measure is one which is likely to give the right answer, rather than, as before, one which is likely to give an answer very close to the right one.

Despite the large differences in definition of error, Henry (1973) has shown that a mathematical development of a dichotomous measurement model is possible which parallels the multiwave path-analytic techniques.

Henry's measurement model assumes that persons in category 1 of a latent dichotomy have some probability q_{11} of giving response 1, and probability q_{12} of giving response 2, where $q_{11} + q_{12} = 1$. Similar but not identical parameters q_{21} and q_{22} characterize persons who are really in category 2. Change in latent state is assumed to happen according to a Markov chain: individuals change state between measurements with probabilities which at any given time depend only on their current state. To identify his model, Henry made the further assumption that response errors occur independently: thus, at any given time, the probability of measurement error, like that of change, depends only upon the latent state.

We should note that these are nontrivial assumptions. Persons who have just changed from one latent state to another may be different from persons who have not changed. They may be unstable, and hence more likely to change again (although here one must determine the point at which instability may be considered measurement error); or they may have changed for a reason, which might render them less likely to revert to their previous state. Likewise, persons who have made erroneous responses may be likely to repeat them, for most of the same reasons that can produce correlated errors in continuous variables. It is not surprising that strong assumptions are needed to extract parameters for measurement error, in the presence of latent change, from a simple two by two by two cross-tabulation. One would like to know how sensitive the technique is to violations of the assumptions, but the algebra connecting the latent processes with the observed crosstabulation becomes quite complex with alternate sets of assumptions, so that it is difficult to dissect the estimation algorithm.

At any rate, we do not know of any discrete-variable models for assessing measurement error, in the presence of change, that can employ weaker assumptions. We are therefore left with the Henry algorithm, and the warning that its Markov change model may be inappropriate for sociological data like ours. The algorithm yields estimates of the probability of error for persons in each category of the dichotomy, at the second of the three time points. We presume that these probabilities would not change very fast, and in particular would not change much over the period of our three measurements. Indeed, as mentioned above, the general three-wave strategy estimates only the reliability at the second measurement, so that one must always assume some constancy of parameters to obtain reliabilities at the first and last measurements.

Detecting Change in the GSS Data

Blalock and Coleman have noted that practical problems arise when one attempts to measure true change over a short time period, if the magnitude of change is much less than that of measurement error (Coleman, 1968, p. 453, n. 13). The techniques we have considered do not avoid these problems. Measurement error may have constant expectation, but it is subject to sampling variability; the problems alluded to by Blalock and Coleman arise when this variability turns out to be as great as the amount of true change. We have some cause for apprehension concerning the GSS reinterviews, then, where all three interviews took place within about three months. (The mean interval between measurements 1 and 2 was 46.9 days in 1973, 46.4 days in 1974; between measurements 2 and 3, 33.3 days in 1973, 32.5 days in 1974.)

As described earlier, one estimates the amount of change between measurements two and three by comparing the strength of the relationship between the first two measurements with that between the first and third. If the (1,3) relationship is actually observed to be stronger than the (1,2) relationship, this strategy breaks down. Such a situation could occur for two reasons, one of which is of some importance. The direction of true change could have reversed sharply, so that (loosely speaking) lots of people changed in each interval, but overall, most of them ended out close to where they had started. This kind of process is implausible, except in very particular circumstances. Furthermore, it is not compatible with the change models assumed by the three-wave techniques, and would not be properly described by their estimation algorithms. Much more likely would be the explanation mentioned above, that fluctuations in measurement error shifted the correlations sufficiently to conceal what change took place between the second and third measurements. These situations must be expected occasionally, and in particular when little true change occurs. They do not contradict the models; but they will produce meaningless parameter estimates, such as stability coefficients greater than unity.

Among thirty-two variables or scales deemed suitable for analysis with Pearson correlations, the mean correlation between measurements one and two was .786, while that between one and three was .783. Furthermore, r_{13} was actually greater than r_{12} for half of the variables (sixteen out of thirty-two; the Heise stability estimates would be greater than one for these sixteen variables.) Among fifty-seven dichotomized variables, the mean proportion of respondents whose answers agreed between first and second measurement was .871; between first and third, .867. Again,

the agreement between one and three was greater than between one and two for a plurality of variables (28 out of 60, with (1,2) greater than (1,3) for 21 varibles and identical proportions for 11 varibles). Measurement error clearly swamped whatever change did take place.

Despite this pattern, we actually do believe that there was change in some variables, even over a mere eleven weeks. Both the reconciliations and a crude analysis of reinterview timing indicate some true change. We attempted to locate it by classifying our variables into three groups: those we were confident would be perfectly or almost perfectly stable; those we expected to be relatively unstable; and an intermediate group. This classification did not help. Even for the changeable and intermediate groups of variables the mean (1,2) relationship exceeded the mean (1,3) relationship by utterly trivial amounts: .003 and .009, respectively, for the correlations, .012 and .013, respectively, for the agreement proportions. Sampling variations in measurement error was evidently far too large for us to make any reasonable estimates of the stability of these variables. (Another possibility, for some varibles, is that "true" change occurred so rapidly as to look like measurement error.)

Thus far we simply know that our variables, in the aggregate, appear stable over the short reinterview period. We need not use the three-wave techniques, but can simply observe that the stabilities are very close or equal to unity over this period. If we liked, we could even choose a few variables whose observed relationships were attenuated over the longer period, and apply the Heise or Henry technique to them. (Of course, the argument of less than perfect stability would be weak statistically given the care with which we would have to select such

variables.) Still, what we have said so far poses no problems for estimation of test-retest reliability, for with no appreciable true change, one can simply use the test-retest correlation. Alas, the real situation is far more complicated.

Correlated Error

One of the most vexing problems in quantitative measurement theory is the possibility that response errors, instead of being random, may be correlated with something else in the measurement model: either with the value of the true variable itself, or simply among themselves at different times. Errors may be correlated with the true variable, for example, because of respondents' desire to appear average or typical, or simply because the scale on which something is measured does not allow for extreme values. Either of these situations forces some reported scores toward the center, so that errors are negatively correlated with true score. Such correlations distort the interval properties of the scale as compared to the true score, so that equal measured intervals do not correspond to equal "true intervals." They may impair or improve such standards of criterion validity as the linearity of relationships with other variables; their effect on measurement of the true variable remains unknown unless one has a very precise definition of the latter.

More serious is the possibility that measurements made repeatedly tend to reproduce the same errors, that is, that errors are correlated among themselves. Errors are reproduced, for instance, if respondents remember their previous answers and try to be consistent. However, there need be no such simple causal connection. Anything other than the true variable--however one defines that--which causes later responses

to be similar to earlier ones leads to correlated error terms. A number of touchy definitions are needed to give substance to the concept of error correlation, as in many aspects of measurement theory. For example, the causal effect of prior variables can be explicitly modelled as such, but could equally be treated as error correlations. Werts, Linn, and Joreskog (1971, pp. 404-406) have remarked on the care that must be taken in distinguishing conceptually between error correlations and various other possibilities.

Methods for numerical estimation of correlations between error terms have been proposed, and in some instances carried out on data thought appropriate to them (Blalock, 1970; Wiley & Wiley, 1974; Wheaton et al. 1977). Although these estimates frequently yield large values for the error correlations, at least where the results are published, they only apply to correlations structured appropriately for the method used in detecting them. It is obvious that the way in which one searches for correlated error limits the kind of correlation one can find. Correlated errors arising from a survey question that is hard to understand could not be statistically detected without multiple indicators, for they would reappear each time the question was asked. Any technique using single indicators would then underestimate the magnitude of error correlation. Multiple-indicator techniques, of course, face the problems of estimation and interpretation we have outlined above.

We shall, however, present some evidence that error correlations are not at all infrequent in test-retest survey data. For reasons that will become clear in the discussion, we do not intend to propose any particular way of dealing with them. The correlated errors appear to arise from the interview format, the length of the questionnaire, the

ordering of the questions, or some combination of these. In both of our three-wave panels, the original or first interview was conducted in person and required over an hour, on average, to complete. The second and third interviews in contrast, were conducted over the phone by a different interviewer, using a much shorter (10 to 15 minutes) questionnaire with the questions ordered differently. If these factors did not influence measurement error, one would expect the (2,3) relationship to be the largest of the three between a third and half of the time. (The (2,3) relationship would be largest a third of the time, by chance, if no true change were involved, only measurement error. If change occurred, the proportion could rise to half or perhaps slightly more, since the 2-3 time interval averaged a little shorter than the 1-2, and was always less than 1-3. We argued in the last section that little or no true change was detectable for most of our variables, since the (1,3) relationship was typically as large as the (1,2).)

Of the 32 sets of test-retest correlations in Appendix 4, r_{23} is the largest in 23 instances. And this understates the difference, for some of the variables are so reliable that a "ceiling effect" obscures differences between the correlations. Among the 23 sets where any of the three correlations is less than .9 (and hence there is room for differences to show themselves), r_{23} is the largest in 19 instances. The pattern among the agreement proportions is the same: the (2,3) agreement is strictly larger than both (1,2) and (1,3) in 39 of 57 instances, over two-thirds. Where one or more proportion is less than .9, the (2,3) relationship is the strongest in 31 of 42 instances, over three-quarters. Considering the means of each measure of agreement, we have:

	Correlations (n=32)	Proportion of agreement (n=57)
(1,2)	.786	.871
(1,3)	.783	.867
(2,3)	.838	.902

Measurements two and three were simply more alike than was either of them with measurement one, which differed in format, length, and ordering. We have already argued, by comparing the (1,2) relationships with the (1,3), that change over these time periods is of negligible magnitude. We therefore believe that the strong (2,3) relationships are not due to the shorter interval between those two measurements. The difference evidently arose from the circumstances of the interview.

The weakening of the test-retest relationship across the changes in interview format, length, and ordering indicates that one or more of these factors affects response error. In the language of path analysis, response errors are correlated among interviews of the same type. This has immediate consequences for the three-wave model in Figure D, which is no longer identified, under any reasonable set of assmptions, if we add correlated error terms between measurements two and three. We feel that, in addition, the implications of these error correlations go much further, affecting all research into test-retest reliability.

Changing various aspects of the way the interview is conducted weakens the test-retest relationship. One's immediate reaction is to say, very well, let us keep the interview situation constant and avoid such disruptions. This would indeed be an effective way of concealing the problem. Behind the scenes, however, whatever extraneous factors affect response error would still operate; errors would still be correlated; and estimates of reliability or stability would be wrong. One could

alternatively try to estimate the size of the error correlations by comparing test-retest relationships within a single type of interview to those across a change; for instance, in the GSS data, by comparing (2,3) relationships with (1,2). Such comparisons would be of little use, however. They would indicate not the total error correlations, but only the extent to which these were disrupted by a particular change in interview medium, length, and ordering. As discussed earlier, the number of possible sources of error correlation is very great. While one can recognize here an opening for a great deal of (rather expensive) experimental work, it is unlikely that results of any practical use will be forthcoming soon. Indeed, as one changes more and more aspects of the interview, attempting to disrupt greater amounts of the error correlations, the necessary assumptions about constancy of random error become quite improbable. As soon as one acknowledges the pervasive presence of nonrandom error, which is consistent across retests but unrelated to what one wants to measure, the prospects for realistically modelling the "true variables" behind the observed ones recede very far. This does not compromise the analysis of survey data in general, for a great deal of insight has been and will continue to be drawn from admittedly imperfect measurements. We do believe that this nonrandom error, which appears to be of magnitude comparable to the random error, invalidates most attempts to purify sociological models from measurement error. Efforts to separate the real from the unwanted in something as complex as a social survey rely, necessarily, on very strong assumptions as to what the unwanted elements will look like: typically, that they will be absolutely unrelated to anything. Perhaps it is fortunate that changes between test and retests in the GSS panels have forced us to give thought to the unlikelihood of these assumptions, and the harmful consequences of making them when they are not appropriate.

Statistical Approaches: Summary

In light of our findings, some of the "practical" uses of sociological test-retest data, in correcting for attenuation and in related methods of latent-variable analysis, appear problematic at best. This is perhaps not a bad thing. The correction for attenuation is a dubious improvement when the possible kinds of measurement error are as varied as they are in sociological surveys. In practice, this "correction" is simply a way of making little correlations into bigger, more interesting correlations, by giving oneself the benefit of every doubt. One assumes that all sources of error are random, that is, that they all act to attenuate the observed correlation. If simple test-retest correlations are used, one also assumes that no true change occurred in the interval. We suspect that it is correct to say that the "true" correlation--to the extent that this can be defined--is usually larger than the observed correlation. Even so, it is hardly good scientific practice to make assumptions for the purpose of exploiting the very imprecision of one's measurements. Researchers inevitably start with the observed correlation, and so find themselves in the awkward position of hoping their measurements are unreliable, so that their findings will be stronger and therefore more interesting.

Our general feeling is that one should be very cautious in claiming to have solved the problems of sociological measurement. Available methods can only be considered hypothetical and exploratory; regrettably, such qualifications have a way of getting lost between the producer and the consumers of research. With the techniques we have surveyed one can write empirical sociology which purports to be free from measurement error. The author of such a study, if competent, will know his findings

for what they are: the implications of a set of assumptions which are not terribly plausible, but which had to be made if any implications at all were to be drawn. This is how science works, and there is nothing wrong with it, so long as one recognizes the extent to which the results were imposed by the research design. These limitations should be acknowledged, however. To assert that one has surmounted the problems of measurement error is simply false. To imply the same (as by phrasing the qualifications in terms of "the usual assumptions" when stating results), is irresponsible; for the hypothetical talk about true scores disarms the natural--and healthy--skepticism of the unsophisticated reader about conclusions drawn from fallible data.

We attempted in the previous section to present evidence that the assumptions required by present techniques for eliminating measurement error are not merely hypothetical, but altogether unacceptable for a variety of survey variables. Although not intended as such, the 1973-74 GSS reinterview panels may be regarded as a quasi-experimental search for response error correlated with the interview situation. We found these kinds of correlated error to be widespread and often substantial. This finding should surprise no one who has thought about the problem. It does reinforce the distinction one must make between methodological sophistication and empirical adequacy when evaluating analytical techniques.

In summary, we have not extracted any magic numbers from the 1973-74 three-wave panels. We do present the means, variances, and correlations in Appendices 3 and 4, for the convenience of people who believe in magic. The three-wave design was chosen for these studies to permit the separation of reliability from stability; we have written at some length to explain why we have failed to carry through the original

purpose. The close spacing of the three waves and the major changes in the interviewing situation blocked our initial attempts to analyze stability and reliability. However, only the conjunction of these "problems" in the design revealed the more serious difficulties inherent to the application of standard latent-variable techniques to real test-retest data.

Conclusion

Our discussion of the GSS test-retest data has had a double focus: the estimation of measurement error and the detection of true change in situations contaminated by measurement error. We have not embraced any of the currently available techniques as suitable for use with our data, so that we seem to be left with the proverbial conclusion that you "can't get there from here." Such pessimism, however, is warranted only if we insist upon a "plug-in" technique, something generally useful that would relieve us of the need to think about the problems arising from measurement error and change. More modest goals are still worth pursuing. The issues of reliability and stability lie at the foundations of empirical sociology, and of social indicators research in particular. Their clarification is a worthy goal, even if they can never be laid entirely to rest.

Study of measurement error in survey data should, we think, be carefully distinguished from the study of simple assumptions about measurement error. We have reached a few conclusions about measurement error and true change as actually found in the GSS test-retest data. For most of our variables, test-retest associations were not weakened during the lapse of a month between second and third measurements. This implies either that they were hardly changing at all, or that change

took the form of more-or-less random fluctuations, statistically indistinguishable from measurement error. Furthermore, we have found that test-retest consistency is higher for demographic variables, which usually refer to objective conditions, than for the more subjective attitudinal and evaluative questions. Among demographics, consistency is naturally greatest for permanent traits. Questions referring to past events or conditions, and questions whose answers could change, show somewhat less consistency.

We have also found that changing the way in which the questions were asked, from an hour-long personal interview to a telephone interview with a rearranged and much shorter questionnaire, reduced test-retest consistency markedly. Since the true scores can hardly have depended upon such matters, it is evident that response errors are related to some aspects of the interview situation. Some response errors are therefore correlated between test and retest.

Our suggestions for future research are varied but prosaic. Present incentives within the profession are quite sufficient to ensure continued work on sophisticated modelling algorithms, including those which incorporate simple assumptions about measurement error. We suggest further work on the fundamentals. Techniques such as validation, reconciliation, and post-interview debriefing attack the problem of response error directly, at the level of the individual respondent. They do not lend themselves to statistical analysis, but can be very helpful in refining question wordings, or simply in determining what a question means to respondents. Development of scales to measure important concepts permits better measurement and some estimate of reliability. Finally, the usefulness of panel methods in measuring change, clarifying causal order, and

untangling the effects of age, cohort, and period is much that further development would undeniably be valuable. A pressing problem in panel studies is the identification of sources of error correlation. An experimental approach to this problem must be taken if one is to do more than study assumptions.

Individually, our conclusions are modest ones. The important point, we feel, is that problems of measurement error and of change are sufficiently fundamental and sufficiently complex that we must attack them piecemeal, converging on them from various directions and with various techniques, but always concentrating on real problems rather than crude idealizations of them.

APPENDIX 1

On the 1972, 1973, 1974, and 1978 General Social Surveys, test/retest data were collected as part of a methodological investigation of reliability and stability. On each survey, a random subsample of respondents on the GSSs were contacted by phone and reasked a selected subset of questions. In 1972 and 1978, the test/retest design consisted of the initial personal interview and a single telephone reinterview. In 1973 and 1974, the test/retest design consisted of the initial personal interview and two waves of telephone interviews for a total of three measurement points. In addition, in 1972 the reinterview subsample was further subdivided into three forms, each with approximately one-third of the reinterviews.

Response rates on the first reinterview ranged between 72 and 82 percent (See Table A-1). Breakoffs and refusals contributed 3 to 6 percent of the nonresponse while inability to contact accounted for 15 to 21 percent. From the more detailed 1978 breakdown of no contacts, it is apparent that having no telephone or not giving a telephone number was the major reason for not being able to make contact. Response rates for the second reinterview were .844 in 1973 and .929 in 1974. Unlike the first reinterview, refusals rather than inability to contact accounted for the majority of nonresponse. Response rates for both the first and second reinterviews (i.e., participated on both) were .619 in 1973 and .670 in 1974.

TABLE A-1

NUMBER OF CASES, AND RESPONSE RATES BY Form/Waves 1972, 1973. 1974, and 1978 Reinterviews

Year	Form/Wave	Attempted Reinterviews	Completed Reinterview	Break-offs	Refusals	No Contact	No Phone/ No Number
1972	All forms	493	.771 (380)	.018 (9)	.043 (21)	.168 (83)	
	Form A	169	(132)				
	Form B	165	(122)				
	Form C	159	(126)				
1973	Wave 1	315	.721 (227)	.013 (4)	.267 (84)		
	Wave 2	231	.844 (195)	--	.156 (36)		
1974	Wave 1	291	.722 (210)	.003 (1)	.062 (18)	.213 (62)	
	Wave 2	210	.929 (195)	--	.043 (9)	.029 (6)	
1978		324	.818 (265)	.003 (1)	.028 (9)	.037 (12)	.114 (37)

The mean interval between the interview and first reinterview was 22.7 days in 1972, 46.9 days in 1973, 46.4 days in 1974, and 33.9 days in 1978. The mean interval between the interview and second reinterview was 80.2 days in 1973 and 78.9 days in 1974.

The 1972 reinterview forms covered 92 items (30 on form A, 40 on form B, and 29 on form C). Because seven questions appeared on more than one form, these total to 99. The 1973 data covered 55 questions on the first reinterview and 44 items on the second reinterview. In 1974, the same 19 items appeared on both the first and second reinterviews.

In 1978, there were 23 items on the reinterview. In addition, in 1972 and 1973 certain questions (13 in 1972 and 4 in 1973) were slated for reconciliation. The respondent's original response was coded on the reinterview form and if the respondent gave a different response on the reinterview, the reinterviewer was instructed to reconcile the different responses. A typical reconciliation probe inquired, "Now in the original interview, the interviewer recorded _____ (READ WHAT WAS RECORDED). Thinking about this for a moment, could you tell me why you think there is a difference between that time and now. (RECORD VERBATIM)."

APPENDIX 2

TEST/RETEST CONSISTENCY[a]

(Percent Agreeing and Pearson Product-Moment Correlation)

	1972	1973	1974	1978
ABDEFECT	.866/.595	.868/.527	--	--
ABHLTH	.885/.539	.940/.612	--	--
ABNOMORE	.875/.739	.861/.721	--	--
ABPOOR	.925/.666	.855/.708	--	--
ABRAPE	.942/.821	.904/.649	--	--
ABSINGLE	.798/.596	.859/.718	--	--
AGE	--	.983/.973	1.000/1.000	--
ATTEND	.960/.919	.945/.920	--	--
BUSING	.855/.491	--	--	--
CAPPUN	.937/.846	--	--	--
CHILDS	1.000/1.000	.961/.923	.976/.950	
CHLDIDEL	.924/.846	--	.833/.667	--
CHLDMORE	.976/.462	--	--	--
CLASS	.845/.700	--	--	--
COLATH	.782/.559	--	--	--
COLCOM	.757/.524	.777/.547	--	--
COLSOC	.748/.460	.761/.524	--	--
COMMUN	--	.806/.610	--	--
CONARMY	--	--	--	.853/.348
CONBUS	--	--	--	.834/.351
CONCLERG	--	--	--	.829/.418
CONEDUC	--	--	--	.842/.263
CONFED	--	--	--	.763/.415
CONFINAN	--	--	--	.870/.229
CONJUDGE	--	--	--	.839/.264
CONLABOR	--	--	--	.712/.377
CONLEGIS	--	--	--	.780/.382
CONMEDIC	--	--	--	.908/.286
CONPRESS	--	--	--	.824/.477
CONSCI	--	--	--	.891/.296
CONTV	--	--	--	.732/.387
COURTS	.982/.894	--	--	--
DEGREE	.977/.951	--	--	--
DRINK	--	--	--	.897/.764
EARNRS	.864/.742	.845/.699	--	--
EDUC	.955/.901	--	-	--
ETHNIC	.976/.939	.945/.887	.968/.937	--
FAIR	.814/.668	--	--	--
FAMILY16	.932/.817	--	--	--
FEPRES	.917/.877	--	--	--
FEWORK	.750/.443	--	.822/.484	--
FINALTER	.746/.462	--	--	--

[a]In 1973 and 1974 the comparison is between the test and first retest. All items are dichotomized and "Don't Knows" are excluded from the analysis.

	1972	1973	1974	1978
FINRELA	.939/.828	.812/.421	--	--
GETAHEAD	-	.892/.306	--	--
GOVAID	--	.814/.565	.833/.625	--
GUNLAW	.865/.661	--	--	--
HAPPY	.866/.483	.872/.427	.842/.314	--
HEALTH	.908/.734	.836/.579	.824/.543	--
HELPFUL	.752/.602	--	--	--
HIT	--	--	--	.878/.723
INCOME	.948/.795	.909/.793	--	--
INCOM16	.947/.876	.867/.671	.787/.535	--
LIBATH	.712/.415	--	--	--
LIBCOM	.796/.557	.854/.699	--	--
LIBSOC	.630/.151	.808/.487	--	--
MADEG	.934/.861	--	--	--
MAEDUC	.913/.873	--	--	--
MARITAL	.976/.937	.978/.948	--	--
MOBILE16	.949/.883	--	--	--
NATAID	--	.738/.567	--	--
NATARMS	--	.871/.537	.786/.506	--
NATCITY	--	.839/.622	.799/.215	--
NATCRIME	--	.941/.305	--	--
NATDRUG	--	.919/.296	--	--
NATEDUC	--	.919/.264	--	--
NATENVIR	--	.926/.312	--	--
NATFARE	--	.802/.605	--	--
NATHEAL	--	.985/.258	--	--
NATSPAC	--	.868/.730	--	--
NEWS	.984/.963	--	--	--
OCC	.930/.859	--	--	--
PADEG	.944/.882	--	.977/.847	--
PAEDUC	.951/.921	--	.950/.855	--
PAOCC16	.963/.914	--	--	--
PAPRES16	.944/.882	--	--	--
PARTYID	.929/.857	.884/.766	--	--
PILL	--	--	.882/.242	--
PREMARSX	.886/.774	--	--	--
PRESTIGE	.896/.778	--	--	--
PRES68	.940/.871	--	--	--
RACDIN	.896/.706			
RACE	1.000/1.000	1.000/1.000	.995/.975	
RACFEW	.958/.439	--	--	--
RACJOB	.972/-.013	--	--	--
RACMAR	.863/.685	--	--	--
RACOBJCT	.952/.722	--	--	--
RACPRES	.885/.578	--	--	--
RACPUSH	.778/.495	--	--	--
RACSCHOL	.966/.657	--	--	--
RACSEG	.777/.499	--	--	--
RADIOHRS	--	--	--	.833/.661
REG16	1.000/1.000	--	--	.985/.969
RELIG	.967/.915	--	--	.943/.872
RES16	.985/.969	.866/.760	.880/.749	--
ROBBRY	--	.974/.237	.947/.579	--

	1972	1973	1974	1978
SATFIN	.828/.588	--	--	--
SATJOB	.781/.561	.853/.716	.721/.445	--
SIBS	.954/.909	.965/.932	--	--
SPKATH	.804/.559	--	--	--
SPKCOM	.764/.587	.848/.685	--	--
SPKSOC	.808/.317	.858/.547	--	--
TEENPILL	--	--	.836/.467	--
TRUST	.813/.643	--	--	--
TVHOURS	--	--	--	.825/.648
UNEMP	--	--	--	.885/.727
USINTL	--	.900/.759	--	--
USWAR	--	.873/.739	--	--
VOTE68	.941/.848	--	--	--
WKSUB	.933/.749	--	--	--
WKSUP	.775/.527	--	--	--
WRKSTAT	.951/.898	.939/.878	.938/.896	--

APPENDIX 3

MEANS AND VARIANCES, 1973-74 GSS INTERVIEWS[a]

Variable	Year	Mean (Variance) Wave 1	Wave 2	Wave 3	N
RES16	73	3.402 (2.521)	3.402 (2.523)	3.474 (2.541)	194
INCOM16	73	2.841 (.655)	2.825 (.666)	2.794 (.633)	189
SIBS	73	4.381 (13.138)	4.309 (12.660)	4.237 (12.306)	194
HRS1	73	40.106 (180.905)	38.365 (181.376)	41.388 (230.360)	85
SPHRS1	73	44.347 (102.528)	43.387 (76.105)	43.920 (107.129)	75
CHILDS	73	2.207 (3.502)	2.319 (3.630)	2.293 (3.727)	188
Tolerance (3 socialist items)	73	2.019 (1.262)	2.223 (1.084)	2.108 (1.366)	157
Tolerance (3 communist items)	73	1.540 (1.525)	1.590 (1.618)	1.665 (1.674)	161
Tolerance (all 6 items)	73	3.667 (4.138)	3.929 (4.666)	3.886 (5.016)	141
FINRELA	73	2.912 (.526)	2.990 (.487)	3.031 (.455)	194
HEALTH	73	2.016 (.748)	1.938 (.708)	1.953 (.715)	192
Abortion (3 strong reasons)	73	3.293 (.546)	3.467 (.889)	3.431 (.849)	167
Abortion (3 weak reasons)	73	4.548 (1.898)	4.548 (1.910)	4.488 (1.973)	166
Abortion (all 6 reasons)	73	7.793 (3.440)	7.987 (4.295)	7.893 (4.351)	150
COMMUN	73	1.853 (.728)	1.972 (.823)	1.994 (.778)	177
SATJOB	73	1.613 (.533)	1.547 (.426)	1.577 (.452)	137

APPENDIX 3 (Continued)

Variable	Year	Mean (Variance) Wave 1	Wave 2	Wave 3	N
ATTEND	73	4.399 (7.450)	4.558 (7.210)	4.793 (6.775)	188
EARNRS	73	1.679 (1.250)	1.751 (1.521)	1.637 (1.566)	193
INCOME	73	7.835 (6.664)	7.688 (6.616)	7.653 (6.594)	176
AGE	73	43.843 (261.217)	43.927 (263.604)	43.953 (263.455)	191
AGEWED	73	21.497 (19.227)	21.604 (23.526)	21.527 (19.382)	169
PARTYID	73	2.678 (3.895)	2.711 (4.117)	2.761 (3.970)	180
HRS1	74	40.828 (141.820)	40.644 (131.930)	41.356 (156.278)	87
SATJOB	74	1.691 (.592)	1.612 (.500)	1.691 (.607)	139
HEALTH	74	2.051 (.760)	1.969 (.772)	1.923 (.721)	195
CHLDIDEL	74	3.006 (2.555)	2.951 (2.640)	2.973 (2.774)	183
CHILDS	74	2.201 (3.239)	2.191 (3.243)	2.216 (3.414)	194
INCOM16	74	2.714 (.593)	2.719 (.800)	2.714 (.823)	192
PAEDUC	74	9.708 (16.271)	9.538 (16.623)	9.677 (16.221)	130
PADEG	74	.610 (.832)	.566 (.766)	.610 (.803)	136
RES16	74	3.361 (2.403)	3.402 (2.449)	3.387 (2.570)	194
AGE	74	44.258 (284.803)	44.247 (284.384)	44.237 (283.528)	194

[a]Figures include only those respondents who answered the question in all three waves of interviewing.

APPENDIX 4

TEST-RETEST CORRELATIONS, 1973-74 GSS

Variable	Year	r_{12}	r_{13}	r_{23}	N
RES16	73	.769	.788	.865	194
INCOM16	73	.731	.684	.837	189
SIBS	73	.905	.959	.933	194
HRS1	73	.866	.836	.797	85
SPHRS1	73	.420	.456	.444	75
CHILDS	73	.922	.939	.920	188
Tolerance (3 socialist items)	73	.655	.604	.775	157
Tolerance (3 communist items)	73	.802	.787	.888	161
Tolerance (all 6 items)	73	.824	.790	.892	141
FINRELA	73	.531	.556	.738	194
HEALTH	73	.677	.738	.776	192
Abortion (3 strong reasons)	73	.753	.751	.862	167
Abortion (3 weak reasons)	73	.831	.860	.914	166
Abortion (all 6 reasons)	73	.865	.875	.917	150
COMMUN	73	.618	.708	.759	177
SATJOB	73	.633	.428	.499	137
ATTEND	73	.944	.909	.949	188
EARNRS	73	.814	.765	.852	193
INCOME	73	.878	.860	.934	176
AGE	73	.9988	.9987	.9998	191
AGEWED	73	.946	.982	.949	169
PARTYID	73	.888	.842	.880	180

APPENDIX 4 (Continued)

Variable	Year	r_{12}	r_{13}	r_{23}	N
HRS1	74	.882	.714	.755	87
SATJOB	74	.563	.553	.741	139
HEALTH	74	.668	.688	.736	195
CHLDIDEL	74	.546	.572	.706	183
CHILDS	74	.981	.973	.973	194
INCOM16	74	.544	.654	.726	192
PAEDUC	74	.943	.956	.957	130
PADEG	74	.937	.937	.945	136
RES16	74	.814	.827	.884	194
AGE	74	.9995	.9996	.9998	194

BIBLIOGRAPHY

Achen, Christopher H.
1975 "Mass Political Attitudes and the Survey Response." American Political Science Reveiw, 69 (December).

Andersen, Ronald; Judith Kaspter; Martin R. Frankel, and Associates.
1979 Total Survey Error: Applications to Improve Health Surveys. NORC Series in Social Research. San Francisco: Jossey-Bass.

Asher, Herbert B.
1974 "Some Consequences of Measurement Error in Survey Data," American Journal of Political Science, 45 ( ).

1974 "The Reliability of the Political Efficacy Items," Political Methodology, 1 (Spring).

Blalock, Hubert M., Jr.
1970 "A Causal Approach to Nonrandom Measurement Errors," American Political Science Review, 64.

1970 "Estimating Measurement Error Using Multiple Indicators and Several Points in Time," American Sociological Review, 35 (February).

1971 Causal Models in the Solcial Sciences. Chicago: Aldine.

1972 Social Statistics. New York: McGraw-Hill.

Bohrnstedt, George W.
1968 "Observations on the Measurement of Change," in Sociolgoical Methodology, 1969, edited by Edgar F. Borgatta. San Francisco: Jossey-Bass.

1969 "Reliability and Validity Assessment in Attitude Measurement," in Attitude Measurement, edited by Gene F. Summers. Chicago: Rand McNally.

Bradburn, Norman M. and Seymour Sudman et al.
1979 Improving Interview Methods and Questionnaire Design. San Francisco: Jossey-Bass.

Coleman, James S.
1968 "The Mathematical Study of Change, " in Methodology in Social Research, edited by H. M. Blalock, Jr. and A. B. Block. New York: McGraw-Hill, pp. 428-478.

Converse, Philip E. and Gregory B. Markus.
1979 "Plus ca Change ...: The New CPS Election Study Panel," American Political Science Review, 73 (March).

Costner, Herbert L.
1969 "Theory, Deduction, and Rules of Correspondence," American Journal of Sociology, 75 (September).

Cronbach, Lee J.
1970 Essentials of Psychological Testing, 3rd edition. New York: Harper and Row.

Cronbach, Lee J.; Goldine C. Gleser; Harinder Nanda; and Nageswari Rajaratnam.
1972 The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley and Sons.

Cureton, Edward E.
1968 "Psychometrics," in International Encyclopedia of the Social Sciences, edited by David L. Sills. New York: Macmillan and The Free Press.

Curtis, Richard F. and Elton F. Jackson.
1962 "Multiple Indicators in Survey Research," American Journal of Sociology, 68 (September)

Davis, James A.
1963 "Panel Analysis: Techniques and Concepts in the Interpretation of Repeated Measurements," unpublished paper. Chicago: National Opinion Research Center.

Dreyer, Edward C.
1973 "Change and Stability in Party Identifications," Journal of Politics, 35.

Erikson, Robert S.
1978 "Analyzing One Variable-Three Wave Panel Data: A Comparison of Two Models," Political Methodology, 5, No. 2.

Ferber, Robert; John Forsythe; Harold W. Guthrie; and E. Scott Maynes.
1969 "Validations of a National Survey of Consumer Financial Characteristics-Savings Accounts," Review of Economics and Statistics, 51 (November).

Heise, David R.
1969 "Separating Reliability and Stability in Test-Rest Correlations," American Sociological Review, 34 (February).

1970 "Comment on 'The Estimation of Measurement Error in Panel Data,'" American Sociological Review, 35, p. 117.

Henerson, Marlene E.; Lynn Lyons Morris; and Carol Taylor Fitz-Gibbon.
1978 How to Measure Attitudes. Beverly Hills, Calif.: Sage Publications.

Henry, Neil W.
1973 "Measurement Models for Continuous and Discrete Variables," in Structural Equation Models in the Social Sciences. edited by A. S. Goldberger and O. D. Duncan. New York: Seminar Press, pp. 51-67.

Keisler, Charles; Barry E. Collins, and Norman Miller.
1969 Attitude Change: A Critical Analysis of Theoretical Approaches. New York: John Wiley.

Kendall, Patricia.
1954 Conflicts and Mood: Factors Affecting Stability of Response. Glencoe, Ill.: The Free Press.

Lord, Frederick M. and Melvin R. Novick.
1968 Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.

Maccoby, Eleanor E.
1956 "Pittfalls in the Analysis of Panel Data: A Research Note on Some Technical Aspects of Voting," American Journal of Sociology, 61, January.

McCullough, B. Claire.
1978 "Effects of Variables Using Panel Data: A Review of Techniques," Public Opinion Quarterly,

McPherson, J. Miller; Susan Welch; and Cal Clark.
1977 "The Stability and Reliability of Political Efficacy: Using Path Analysis to Test Alternative Models," American Political Science Review, 71, ( ).

Nunnally, Jum C. and William H. Wilson.
1975 "Validity, Reliability, and Special Problems of Measurement in Evaluation Research," in Handbook of Evaluation Research, edited by Elmer L. Struening and Marcia Guttentag. Beverly Hills, Calif.: Sage Publications.

Robinson, John P.; Robert Athanasiou; and Kendra B. Head.
1969 Measures of Occupational Attitudes and Occupational Characteristics. Ann Arbor, Mich.: Institute for Social Research.

Robinson, John P.; Jerrold G. Rusk; and Kendra B. Head.
1968 Measures of Political Attitudes. Ann Arbor, Mich.: Institute for Social Research.

Robinson, John P. and Philip R. Shaver.
1973 Measures of Social Psychological Atttidues. Ann Arbor, Mich.: Institute for Social Research.

Schnaiberg, Allan, and Michael Amer.
1972 "Measurement Evaluation Obstacles in Sociological Surveys: A Grounded Reassessment." Paper presented to the American Sociological Association, New Orleans.

Selltiz, Claire; Lawrence S. Wrightsman; and Stuart W. Cook.
1976 Research Methods in Social Relations. New York: Holt, Rinehart and Winston.

Smith, Tom W.
1979 "Can We Have Any Conifence in Confidence? Revisited." GSS Technical Report No. 11. Chicago: National Opinion Research Center.

Sudman, Seymour and Norman M. Bradburn.
1974 Response Effects in Surveys: A Review and Synthesis. NORC Monographs in Social Research. Chicago: Aldine.

Traugott, Michael W. and John P. Katosh.
1979 "Assessing Response Validity in National Surveys of Voting Behavior." Paper presented to the American Association for Public Opinion Research, Buck Hill Falls, Pa., June.

U.S. Bureau of the Census.
1964 Accuracy of Data on Population Characteristics as Measured by Reinterviews. Series ER60 No. 4. Washington, D.C.: Government Printing Office.

1968 The Current Population Survey Reinterview Program, January 1961 through December 1966, Technical Paper No. 19. Washington, D.C.: Government Printing Office.

1972 Effects of Different Reinterview Techniques on Estimates of Simple Response Variance. Series ER60, No. 11. Washington, D.C.: Government Printing Office.

Webb, Eugene J.; Donald T. Campbell; Richard D. Schwartz; and Lee Secgrest.
1966 Unobtrusive Measure: Nonreactive Research in the Social Sciences. Chicago: Rand McNally.

Werts, Charles E.; Karl G. Joreskog; and Robert L. Linn.
1971 "Comment on 'The Estimation of Measurement Error in Panel Data'," American Sociological Review, 36, ( ).

Werts, Charles E. and Robert L. Linn.
1970 "Cautions in Applying Various Procedures for Determining the Reliability and Validity of Multiple-Item Scales," American Sociological Review, 35 (August).

Wheaton, Blair; Bengt Muthen; Duane F. Alwin; and Gene F. Summers.
1976 "Assessing Reliability and Stability in Panel Models," in Sociological Methodology, 1977, edited by David R. Heise. San Francisco: Jossey Bass.

Wiggins, Lee M.
1973 Panel Analysis: Latent Probability Models for Attitude and Behavior Processes. San Francisco: Jossey-Bass.

Wiley, David E. and James A. Wiley.
1970 "Estimating Measurement Error Using Multiple Indicators and Several Points in Time," American Sociological Review, 35 (February).

Wiley, James A. and Mary G. Wiley. "A Note on Correlated Errors in Repeated
1974 Measurements," Sociolgoical Methods and Research, 3 (November).